Downloaded from orbit.dtu.dk on: May 04, 2023

DTU Library

Agent Based Individual Traffic Guidance

Wanscher, Jørgen

Publication date:
2009

Document Version
Early version, also known as pre-print

Link back to DTU Orbit

Citation (APA):
Wanscher, J. (2009). *Agent Based Individual Traffic Guidance*. Technical University of Denmark, Informatics and Mathematical Modelling. IMM-PHD-2007-160

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Agent Based Individual Traffic Guidance

Jørgen Bundgaard Wanscher

Kongens Lyngby 2007
IMM-PHD-2007-160

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Summary

This thesis investigates the possibilities in applying Operations Research (OR) to autonomous vehicular traffic. The explicit difference to most other research today is that we presume that an agent is present in every vehicle – hence Agent Based Individual Traffic guidance (ABIT).

The next evolutionary step for the in-vehicle route planners is the introduction of two-way communication. We presume that the agent is capable of exactly this. Based on this presumption we discuss the possibilities and define a taxonomy and use this to discuss the ABIT system.

Based on a set of scenarios we conclude that the system can be divided into two separate constituents. The immediate dispersion, which is used for small areas and quick response, and the individual alleviation, which considers the longer distance decision support.

Both of these require intricate models and cost functions which at the beginning of the project were not previously considered. We define a special inseparable cost function and develop a solution complex capable of using this cost function.

In relation to calibration and estimation of statitical models used for dynamic route guidance we worked with generating random number sequences. During this work we made significant findings related to random numbers.

Resumé

Denne afhandling undersøger mulighederne i at benytte operations analyse i forbindelse med biltrafik. Den specifikke forskel i forhold til anden forskning i dag er, at vi antager at der forefindes en agent i hver bil – deraf Agent Baseret Individuel Trafikvejledning (ABIT)

Den sidste tekniske udvikling for ruteplanlæggere i dag er tilføjelsen af to-vejs kommunikation. Vi antager, at agenten er i stand til netop dette. Baseret på dette diskuterer vi mulighederne i operations analyse og definerer en taxonomi, der muliggør den efterfølgende diskussion af ABIT-systemet.

Efter gennemgang af et sæt scenarier konkluderer vi, at systemet kan opdeles i to adskillelige dele. Umiddelbar spredning, der arbejder med små områder og på kort sigt, og individuel afhjælpning, der betragter større områder og på længere sigt.

Begge disse virkefelter kræver specielle modeller og omkostningsfunktioner, vi ikke har fundet tidligere litteratur om. Vi definerer en speciel ikke separabel omkostningsfunktion og udvikler en løser, der kan håndtere den øgede kompleksitet og derved drage nytte af de mere realistiske løsninger.

I relation til udviklingen af omkostningsfunktionen og trafikmodelering arbejdede vi med generering af tilfældige tal. Her har vi fundet nogle vigtige retningslinier for brug af disse i forbindelse med statistik generelt.

Preface

This thesis was prepared at the department for Informatics and Mathematical Modeling, the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

The thesis deals with different aspects of mathematical modeling of traffic and online application of operational research. The main focus is on discussing the necessary prerequisites and technological advanced for actual realization of Agent Based Individual Traffic guidance (ABIT).

The thesis consists of a summary report and a collection of 5 research papers written during the period 2003–2006, two published as conference papers and one as a technical report.

Lyngby, August 2007

Jœrgen Bundgaard Wanscher

Papers included in the thesis

- [A] Agent Based Individual Traffic guidance presented on Aalborg Trafikdage 2004
- [B] Agent Based Individual Traffic guidance: Perspectives presented on Aalborg Trafikdage 2006
- [C] ABIT measuring
- [D] Quasi-Newton Method for TAP with inseparable cost function
- [E] Further results
- [F] Drawing a random number published as technical report

Acknowledgments

Statens Teknisk-videnskabelige Forskningsråd is thanked for providing the funds for this project.

Jens Clausen, Professor and Jesper Larsen, Associate Professor at Operations Research at the department of Informatics and Mathematical Modelling, Danish Technical University are thanked for their commitment and enthusiastic effort as supervisors on this project. Furthermore the first visionairy document on ABIT, which was the application to Statens Teknisk-Videnskabelige Forskningsråd was written by them.

I thank David Ryan, Professor at Engineering Science, The University of Auckland, New Zealand for granting me an interesting stay and many fruitfull discussions. Lecturer Judith Wang, Department of Civil and Environmental Engineering, The University of Auckland, New Zealand is also thanked for the comments and discussions during my stay.

x

Contents

Summary **i**

Resumé **iii**

Preface **v**

Papers included in the thesis **vii**

Acknowledgments **ix**

1 Introduction **1**

1.1 Traffic Science: Traffic Assignment 5

1.2 Operations Research 7

1.3 Road Map to Thesis 8

2 Concepts and definitions **9**

2.1 Terminology and notation 9

2.2 Agent Based Individual Traffic guidance 13
2.3 Scenarios 17
2.4 Complex traffic systems 25
2.5 System 27
2.6 Summary 27

3 History **29**

3.1 Theoretical evolution 29
3.2 Solution methods 46
3.3 Application of mathematical theory 49
3.4 Applied practice 50
3.5 Summary 53

4 Theory **55**

4.1 Our model 55
4.2 Basic Solution Method 57
4.3 Time and space 59
4.4 Forecasting 61

5 Papers included in the thesis **63**

5.1 Agent Based Individual Traffic guidance 63
5.2 ABIT measuring 64
5.3 Agent Based Individual Traffic guidance 2 65
5.4 Quasi-Newton Method for TAP with inseparable cost function . 65

5.5 Further results 66

5.6 Drawing a random number 66

6 Conclusion **67**

A Agent Based Individual Traffic guidance **77**

A.1 Traffic Today 78

A.2 The Future for Road Based Traffic 82

A.3 Disruptions 83

A.4 Further possibilities 91

A.5 Conclusion 92

B Agent Based Individual Traffic guidance – Second Presentation **95**

B.1 Overview 96

B.2 Agent Based Individual Traffic guidance 96

B.3 The Future for Road Based Traffic 99

B.4 Theoretical Requirements 100

B.5 Computing Strength 103

B.6 Vehicle Information Systems 103

B.7 Information gathering and processing 104

B.8 Conclusion 105

C ABIT Measuring **109**

C.1 Introduction 109

C.2 Measurable Characteristics 111

C.3 Aggregation 113

C.4 Objective 117

C.5 Conclusion 123

D Quasi Newton Method used for TAP with inseparable cost function 127

D.1 Introduction and background 128

D.2 Agent Based Individual Traffic guidance 128

D.3 A More Precise Cost Function 129

D.4 The solver 132

D.5 Quasi-Newton Projection 133

D.6 Experimental results 135

D.7 Conclusion 137

E Further Results 143

E.1 Path generation 143

F Drawing a random number 149

F.1 Introduction and background 150

F.2 Theoretical framework 151

F.3 Drawing a single number 152

F.4 Multidimensional draws 158

F.5 Obtaining the results 160

F.6 Runtime analysis 161
F.7 Correlation 164
F.8 Conclusion 169
F.9 The pseudorandom draw 171
F.10 The standard Halton draw 172
F.11 The leaped Halton draw 173
F.12 The scrambled Halton draw 174

G Acronyms **177**

Chapter 1

Introduction

The commuting of people is inevitable. We frequently require to travel from one location to another. Many need to travel to their work, go the grocery store or simply travel to or for leisure. In every case the method of transportation must be chosen among several alternatives. These include walking, cycling, public transport or private car.

Why and how we choose among the different alternatives is a question within psychology. In this thesis it is sufficient to see the consequences of the choice. As shown in table 1.1 every year the number of vehicles on the roads increase. This is mainly caused by the general perception of ease of use. It is much easier to go by car than to wait for busses or trains. In Denmark the recent significant

Road	Count location	1993	1997	2001	2005	increase	percent
E20	Taastrup	59.200	66.600	73.900	83.300	24.100	41%
E47/E55	Nærum	56.400	61.700	70.700	73.800	17.400	31%
E47/E55	Husum	52.100	64.900	71.800	75.400	23.300	45%
E20/E47/E55	Hundige	70.300	82.700	91.300	102.300	32.000	46%
H.C. Andersens Blvd	Langebro	54.400	62.600	62.400	60.500	6.100	11%
Vejlands Alle	Sjællandsbro	32.100	32.800	49.900	52.500	20.400	64%
sum		324.500	371.300	420.000	447.800	123.300	38%

These numbers are the average daily counts for vehicles passing the count locations. Source and copyright: Statistics Denmark

Figure 1.1: Vehicle counts on main roads in Copenhagen

increases in public transport pricing could be a major contributor to the increase in private vehicular transport.

The increase in private commuters does not come without a price. Rush hour, formerly a problem only in really big cities, is spreading and intensifying in every major city around the world. The time not spent, but wasted, waiting in rush hour queuing is vast and the socioeconomic impact is devastating. Any change in traversal time on E20 at Hundige during rush hour affects at least 60.000 vehicles. At present the section is ligthly congested and within a relatively short timespan the traversal time can be expected to increase by at least 5 minutes. Combining these numbers indicates that this probably small estimate on traveltime increase for the individual results in no less than 300.000 minutes (5.000 hours or 625 8-hour working days) are lost just in transit. The regional traffic survey [Pedersen, 2004] states that at that time 30 million man hours equivalent of 5.7 billion DKR (approximately 1 billion USD) are wasted around Copenhagen on a yearly basis.

The ecological consequences are truly frightening and every major city has for a long time been working to reduce the direct and indirect impact of the increased traffic. Both air and noise pollution is suspected of causing harm to humans. The exhaust from the vehicles contributes is a significant part of the total CO_2 outlet. In Denmark the vehicular traffic caused 19% of the total outlet [Flagstad, 2006].

It has long been realized that the distribution of information is essential to increase the efficiency of the roads. Obtaining the information has been and is still done manually either by first person traffic reports to central information offices or by surveillance.

Traffic radio broadcast was the first of the generally available means of distributing the essential information. The increased information given to the individual drivers has historically shown to improve both through-put and safety. The drawback, however, is that it is impossible to know if people get the information and if so, how they react to it.

Over the last decade Dynamic Traffic Signs (DTS) or Variable Message Signs (VMS) has won increasing popularity as these enable the display of the information where and when it is needed. Furthermore it does not require anything from the car radio. During the large beltwey 3 expansion in Copenhagen VMS is deployed at several places to warn drivers of variable speedlimits, upcoming queues or expected travelling times.

Concurrently the in-car Global Positioning System (GPS) have increased the static information on the network for all travelers and this is beneficial for

traffic safety and minimizing the kilometers traveled. The drawback of GPS, however, is that the information they present is static. The proposed routes are indifferent to the actual state of the road infrastructure.

Recently, Traffic Message Channel (TMC) [Forum, 2004] was established in Europe as a non-profit forum for developing and deploying a new and more efficient distribution of traffic information. The crucial difference in TMC enabled GPS units is that the information for proposing routes is no longer static. The information broadcast in TMC is idealy all information relating to changed network conditions. Road works, lane closures, variable speed limits and queuing can be transmitted to the GPS routeplanning devices. The devices can then incorporated the dynamic information in the route generation.

The evolution of decision support for route planning has evolved in several stages. Some of these are:

- Traffic Radio: General broadcast of dynamic information
- GPS: Route planning assistance based on static information
- DTS: Visual display of significant information
- TMC: One-way information directly to the vehicle driver

All through these stages, the real time status information is centrally collected and then distributed so that each and every driver can use it individually.

The intent of this thesis is to consider further evolution of the decision support systems in traffic.

Agent Based Individual Traffic guidance (ABIT), the subject of this thesis, is based on the assumption that vehicles are equipped with two way communication. The vehicles are capable of receiving information more complicated than TMC, but are also able to transmit information back to a Central Information System (CIS) or simply distribute it locally as in Life Warning System (LIWAS), [Bronsted et al., 2005].

The terms Ambient or Pervasive combined with Computing or Intelligence, as in Ambient Computing (AC) or Pervasive Intelligence (PI) are well established terms. These refer to systems or deployments where computers are everywhere and connected. The key issues are:

Invisibility The computers are not directly visible.

Construction New possibilities are possible by combining existing components.

Heterogeneity The components can work in different contexts and configurations.

Change Adaptation to changing user needs and underlying technical components.

Scalability Solutions must be considered for almost unlimited contexts.

In our case the computing power is placed in every single vehicle. These computers are capable of communicating both with each other and with a common information network. The system is pervasive as the computers are in every vehicle and their main purpose is to intelligently relay information related to traffic to the driver. We define these kinds of systems as **Pervasive Traffic Intelligence (PTI)**.

The difference to systems like Advanced Traffic Information System (ATIS) or Advanced Traffic Management Systems (ATMS), which would be Traffic Intelligence is that these are not utilizing computers in every vehicle. PTI differs from TMC systems and other one-way systems by the fact that they use two-way communication so that every node in the network can generate and relay information.

Pervasive Traffic Intelligence (PTI) covers all systems that uses two-way communication to every vehicle or between vehicles to increase the situational awareness of the driver. The LIWAS is PTI even though it does not propose routes as ABIT.

In ABIT the information sent from the vehicles to the CIS includes:

- Current position
- Desired destination
- Real time information, such as traveling speed or surface temperature, on local road infrastructure.

Based on this information it is possible to use far more complicated methods. We develop an enhanced model and a solver capable of efficiently utilizing this model.

The article included in appendix A elaborates further on the intent and expected use of ABIT.

Wardrop's first principle
Under equilibrium conditions traffic arranges itself in a congested network in such a way that no individual trip maker can reduce his path cost by switching routes

Wardrop's second principle
Under social equilibrium conditions traffic should be arranged in congested networks in such a way that the average (or total) travel cost is minimized.

Figure 1.2: Wardrop's principles

The thesis covers the theoretical research and progress in the fields of Traffic Science (TS) and Operations Research (OR). It will be clear that these two fields can through ABIT increase both the volume of available information and the application of this essential knowledge.

The impact of ABIT will be extended flexibility to accommodate disruptions in the day to day traffic streams.

1.1 Traffic Science: Traffic Assignment

The field of Traffic Science (TS) covers a wide range of research topics. The common theme of these is the real world application of theory to traffic related problems. This could be estimating the increase in throughput by expanding a specific section. Forecasting the impact of increased service in public transport on the congestion level is also TS.

In the early days of TS, processing and solution time effectively limited the possible applications of TS to situations that could be analyzed and precalculated over a fairly long period of time. This primarily meant that the research in the field could only be applied to long term infrastructural planning.

Even though the costs of applying TS was high, the cost and impact of the choices ensured that much effort was put into making qualified decisions. Since Wardrop stated his first and second principle shown in figure 1.2 the main goal has been to find the user equilibrium defined by the first principle. This is known as the Traffic Assignment Problem (TAP) or Dynamic Traffic Assignment (DTA). The DTA was most likely first mentioned in [Yagar, 1971], but it is

unclear when the TAP was defined.

The difference between the principles is also addressed in game theory. The prisoners dilemma is a problem example from game theory where the two distinct optimality definitions are considered.

These problems aim at finding the flow distribution on the individual sections that result in a state satisfying Wardrop's first principle. By accepting Wardrop's first principle as valid, a solution to a given TAP or DTA instance will thus yield insights into the flow requirements of the road network.

Deciding which roads or movements in the road infrastructure to augment or redesign was done by solving the TAP for each and every augmentation or change. When considering improving a regular intersection with four different roads the possible augmentations could be adding a turning lane, making green periods for turning, extending the lane width of the sections or even illegalizing some turning movements. These four possibilities would yield 16 different scenarios unless some of the augmentations were contradictory. The results were then used in the often political decision process on which scenario, if any, should be implemented.

The size and complexity of the required TAPs were completely intractable for computing until the approach proposed in [Frank and Wolfe, 1956]. This method which will be described in detail in chapter 3 on page 29 was the first promising approach to the TAP.

For the first time it was viable to investigate TAP instances with fairly complicated networks and get solutions that were usable. As the processing power of the computer increased through the later half of the 20^{th} century, the models and the sizes of the networks manageable by the Frank-Wolfe method increased.

Many other problems in TS have evolved since then and the field contains almost anything from model calibration to prediction and forecasting of anything from pedestrians to heavy haul vehicles. The common point to all problems is that they are related to commuting of people or goods.

In this thesis the primary interest is the TAP, DTA and the research concerning distribution of vehicular traffic.

1.2 Operations Research

Operations Research (OR) is as Institute for Operations Research and the Management Sciences (INFORMS) formulates it to the public: "The Science of Better" [INFORMS, 2004]. OR is the scientific field of decision support. Optimizing the output from a scarce set of resources or minimizing resource consumption given some demand are classical problems in OR. Through detailed analysis a mathematical model is constructed. This model is then solved to optimality by mathematical or algorithmic methods or approximated through heuristics.

However OR also covers the softer decision problems related to facilitation and adaptation of work processes.

Operations Research was first recognized as a separate field during the World War II. It was used by the warring parties to make better decisions. The combination of theoreticians and practical war makers lead to more qualified decisions.

One of the more interesting decisions is mentioned in [Postrel, 2004]. It shows that viewing a problem from another angle with new eyes can give good alternatives. When Royal Air Force (RAF) decided to improve the bomber planes they made ground crew note all the damage that was on the returning planes and based on that proposed that the most damaged areas were reinforced. The theoreticians quickly opposed this proposition by suggesting that all the areas that were not damaged should be reinforced. The reason for this contradictory statement was that RAF only saw the returning planes, not the ones that were actually lost. As all the returning planes had suffered damage at the reported areas the ones that did not return might have been damaged elsewhere.

In this sense OR is a theoretical angle that through a model uses mathematics and computers to produce solutions or information.

Through seven decades scientist in the field has continously produced better and more efficient methods. Today, Variational Inequality Problems (VIP), Mixed Complementarity Problems (MCP) and Integer Programming (IP), are just a few of the model classifications. For these models a host of different methods have been developed, proven, tested, benchmarked, discarded and reconsidered. The combination of application of advanced mathematics and the advancements within computer power has enabled OR to solve problems that only few years earlier were deemed impossible or intractable to solve. The huge potential in OR has widened the scope of OR to an excessive amount of subfields each concerning significantly different problems and solution methods. Examples of subfields are Linear Programming (LP), Quadratic Programming (QP), Non-Linear Programming (NLP) or VIP. Problems like crew scheduling, packing,

vehicle routing or traffic assignment have been addressed within several different subfields each. A problem is therefore not confined to only one type of model. Different researchers have proposed different models for the same problems and it is then interesting to see the strength of each model compared with it's tractability. As we shall see later in section 3.2 on page 46 on solution methods, the TAP has been modelled in several different ways.

In this thesis traffic assignment is obviously the key problem. In chapter 3 we give an overview of the different model types that has previously been and are still used to describe the problem.

1.3 Road Map to Thesis

This thesis is divided into 6 chapters and an appendix with relevant papers. First, this introduction has set the context of the thesis. The following chapters on concepts and history will be a more detailed coverage of the research in both fields that this thesis is based on. Chapter 4 will then cover the theoretical issues considered and developed for ABIT. The last chapter before the conclusion will briefly cover the essentials of each of the included papers.

Chapter 2

Concepts and definitions

The intention of this chapter is to define a framework for discussing different situations in the Agent Based Individual Traffic guidance (ABIT) system. First we describe terminology. Following this the ABIT system is used to exemplify normal operation of the system. After this we discuss the term disruption in the given context and the different types of such.

Finally a few scenarios are depicted.

2.1 Terminology and notation

This section covers terminology and notation used through out this thesis. The amount of previous literature is vast and the terminology and notation used in the past differs considerably. This is specially difficult when different research and application fields are considered.

We include this section on the essential words and mathematical formulations used here to ease the understanding of this thesis. The list of acronyms is included last as the appendix G on page 177.

2.1.1 Terminology

Intersection is a junction between roads. It is a general term for both light controlled junctions and uncontrolled junctions.

Section is a confined road with no intersections. In other words it is a piece of road where vehicles at most can change to the opposite direction.

Area is a collection of physically connected sections or intersections. Areas can overlap. An area may be a single section.

Subregion is used for a collection of areas that are physically connected. Subregions are disjoint except for connecting sections and thus cannot logically overlap.

Adjacent subregions is used for subregions that are physically connected by a section or intersection.

Region is a collection of subregions. As with subregions any two regions are disjoint except for connecting sections. The requirement on physical connectedness is dropped and a region can thus contain subregions that are not adjacent to any other subregion within the region. A set of several regions will be called a super region.

Connecting section is a section or intersection that connects two subregions or regions.

Throughput and load are used interchangeably as the current level of traffic in an area, on a section or through an intersection.

Capacity is the amount of traffic that can flow through an area. The capacity of a section can be defined or estimated through well calibrated models. The capacity of a larger area is, however, much more complicated to estimate.

Infrastructure is any part of the transportation system publicly accessible to vehicles.

Path is the logical composition of connected sections and intersections by which a vehicle can move legally from one position to another in the network

Route is the path actually traversed by a vehicle. Thus a route is always related to a specific vehicle as opposed to a path, which is related to the underlying network.

Figure 2.1 illustrates some of the above terms by an example network.

There is a potential difficulty in the regional description above. Long sections (highways, tunnels, etc.) may pass through regions without being part of those regions. The section passes through the region physically, but is not a logical part of it. Furthermore it has to be distinctively mentioned that the subregions cover all sections – there cannot be a section that is not part of at least one subregion. A section is part of two subregions if and only if it is a connecting section.

2.1.2 Notation

We define the following notation:

x_o	Generic origin of trip
x_d	Generic destination of trip
$\overrightarrow{x_a x_b}$	Section from x_a to x_b
$\overrightarrow{x_a x_b x_c}$	Exact path from x_a to x_c via x_b
$p(x_a, x_b)$	Any non-cyclic path from x_a to x_b
$p(x_a, x_b, x_c)$	Any non-cyclic path from x_a to x_c via x_b
$P(x_a, x_b)$	All non-cyclic paths from x_a to x_b

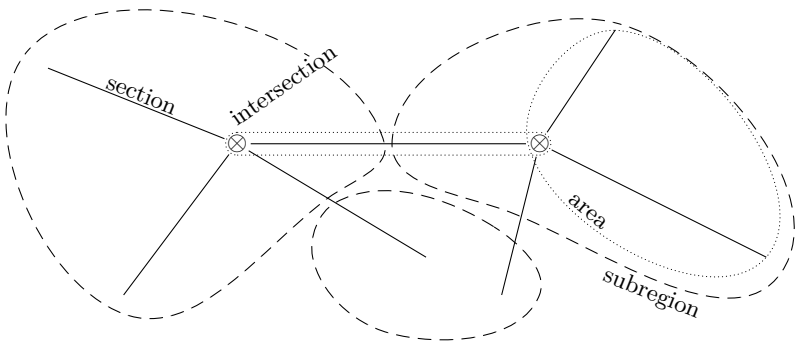


Figure 2.1: Example network with descriptions

We use $\hat{\cdot}$ to indicate optimality, $\lceil\cdot\rceil$ to indicate the capacity, and $|\cdot|$ to indicate the load.

$\widehat{p}(x_a, x_b)$	Set of optimal paths from x_a to x_b
$\widehat{p}(x_a, x_b, x_c)$	Set of optimal paths from x_a to x_c via x_b
$\lceil x_a \rceil$	Capacity of intersection x_a
$\lceil \overrightarrow{x_a x_b} \rceil$	Capacity of the section $\overrightarrow{x_a x_b}$
$\lceil P(x_a, x_b) \rceil$	Combined capacity of all paths from x_a to x_b
$\lvert x_a \rvert$	The current flow through or load on intersection x_a
$\lvert \overrightarrow{x_a x_b} \rvert$	Current flow on the section from x_a to x_b
$\lvert P(x_a, x_b) \rvert$	Total flow from x_a to x_b on all paths

The usage of combined and total in relation to capacity and flow is deliberate. Flow on several paths can easily be added and thus a good aggregate measure is the total. This is not the case with capacity. Simply adding the capacity of each individual path makes no sense as the capacity is bound to each intersection or section. Finding the capacity of a set of paths from an origin to a destination is thus not automatically the total of all the considered paths. The capacity each path must be combined with the underlying network to yeild a usable aggregate. Based on the network in figure 2.2 the following examples illustrate the notation:

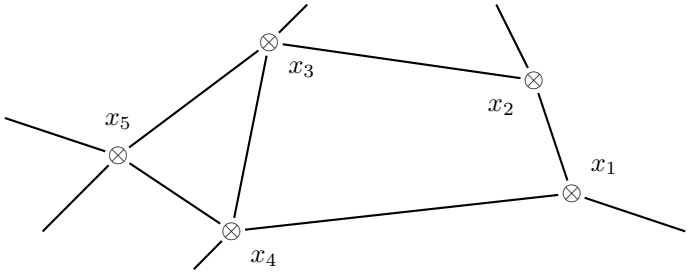


Figure 2.2: Example network for notation

$$P(x_1, x_2) = \{\overrightarrow{x_1 x_2}, \overrightarrow{x_1 x_4 x_3 x_2}, \overrightarrow{x_1 x_4 x_5 x_3 x_2}\}$$
$$P(x_1, x_3, x_2) = \{\overrightarrow{x_1 x_4 x_3 x_2}, \overrightarrow{x_1 x_4 x_5 x_3 x_2}\}$$

If we assume a pricing function f_p we can write $\widehat{p}(x_a, x_b)$ as:

$$\widehat{p}(x_a, x_b) = \{p \in P(x_a, x_b) | f_p(p) \leq f_p(u), \forall u \in P(x_a, x_b)\}$$

This is interpreted as a path p in the set of all paths from x_a to x_b, $P(x_a, x_b)$, which has less or equal price as all individual paths u in the set. It is important to realize that this mathematical definition allows for several different paths $p_1, p_2, p_3, \ldots, p_j$ to be optimal as long as they all have equal costs $f(p_1) = f(p_2) = f(p_3) = \ldots = f(p_j)$.

We will now turn the attention specifically towards ABIT.

2.2 Agent Based Individual Traffic guidance

This section covers the basic description of ABIT and the potential in the system.

2.2.1 Normal operation

Under normal operation the driver simply enters his vehicle and informs the system of the intended destination. The system will then based on the vehicles current position and available routing information guide the driver along the expected best[1] route. If several routes are equivalently rated the system may let the driver decide. The system simply acts as decision support when the driver is to select his route.

At any time during the trip the system may change the route based on local or remote information. As the destination is known, the driver can receive remaining Expected Traveling Distance (ETD) and Expected Traveling Time (ETT).

At the destination the driver stops the vehicle and logically leaves the system.

From the systems perspective the situation is equally simple. When the driver announces the destination the system selects the above described path and continuously keeps the agent updated such that ETD and ETT can be estimated. On completion of the trip the vehicle is no longer part of the system.

[1]The word best is chosen as it opens up for a broader sense of optimality (shortest, fastest, cheapest)

2.2.2 User imposed deviation

As above the driver simply enters the vehicle and designates the destination. The difference is that the driver does not follow the proposed path. Due to the two-way communication of ABIT this deviation will be known immediately by both the agent situated in the vehicle and the system. A new path is thus proposed to the driver by the system and the system returns to normal operation.

2.2.3 Disruptions

We begin by defining the characteristics of a disruption in ABIT. Then we discuss the different types of disruptions.

In any case it is the point of view that decides when an event is a disruption or not. The degree of dynamism in the model sets the limit. If the model is highly static even slight deviations will disrupt the solution. On the other hand highly dynamic models are too complex to compare different possible solutions or even find a solution - optimal or not.

In ABIT, a change in the expected situation of a single vehicle is not a necessarily disruption. If a vehicle slows down or accelerates beyond the expected behavior it does not automatically constitute a disruption. Any situation that unexpectedly interferes in the traffic flow capabilities of a section is a disruption. This distinction is important as it is virtually impossible to predict the exact speed of an individual driver. We thus alleviate this and allow for greater focus on Disruption Management (DM) by considering the overall flow of a section and not the individual vehicles currently traversing a section. This mesoscopic approach is widely adopted in the traffic simulation and forecasting software of today.

The previous distinction divides deviations into two - disruptions and those imposed by users. The problem is that the group of disruptions is far too coarse grained to be used consistently. Therefore, we propose a characterization scheme for a disruption in traffic situations.

2.2.3.1 Temporal and spatial locality

When assessing a disruption, temporal and spatial locality can be used for classification.

Temporal locality is the time characteristic description of the disruption. It can be **short**, when the disruption is quickly recovered. If the disruption is severe it may take a long time to recover, thus the disruption exhibits **long** temporal locality. In between short and long temporal locality, **medium** is placed as further subdivision. The disruption might, in the extreme, never be recovered, constituting **permanent** temporal locality. These distinctions only cover coherent disruptions, and as the system is viewed over time further distinctions could be useful.

Periodical disruptions occur following a possibly non determinable schedule, but are bound by the same duration at each occurrence. Intermittent disruptions are not bound in either duration or schedule and can thus be treated as different disruptions. If the schedule of a periodical disruption is known, the disruption might be assumed to be advance knowledge.

Once again, the dynamics of the ABIT system requires a clearer definition of advance knowledge. **Advance knowledge** is what is known when the problem is initially solved. The problem here is the definition of initially. This system is not bound by working hours or daily schedules and the initial solution is simply the current state of the system and the routing scheme. Advance knowledge is then information about a disruption known sufficiently long before it occurs to ensure smooth and possibly optimal alleviation.

By simplifying intermittent disruptions the distinctions needed to classify the temporal locality of a disruption is reduced to: short, medium, long or permanent and periodical. Any disruption can further more be advance knowledge.

Spatial locality is the physical area characteristic and can be divided into: a) **sectional**, confined to a single section or intersection; b) **multisectional**, influencing several sections or intersections. c) **subregional**, affecting a subregion; d) **regional**, a region is involved; e) **super regional**, influencing on an entire super region; f) **global**, the entire system is disrupted. Classification b) through e) are used to reflect the disruption in relation to subdivision of the network described in section 2.1.1.

All of the above classifications define the spatial locality statically. As with temporal locality further distinction might be necessary. We thus introduce **roaming** disruptions that through time change their physical location with or without changing their spatial class.

Examples are given in section 2.3 on page 17 concerning scenarios.

Having dealt with temporal and spatial locality we turn our attention to the second classification scheme.

2.2.3.2 Level of Influence

Disruptions can also be classified by their **Level of Influence (LoI)**. The LoI is a measure of the impact of the disruption on the system. We define 5 levels: ultra-light, light, mediocre, severe and total.

A disruption that is classified as **ultra-light** have almost no impact on the system. This class is the least demanding for the system as the event is locally confined to a section and quickly recovered.

Light disruptions require more from the system. Dispersion of traffic in a local area should suffice to alleviate this kind of disruption. A **mediocre** disruption influence an entire region, requiring dispersion within the region. **Severe** disruptions require alleviation in several regions and a total disruption requires global alleviation.

This scheme also allows disruptions to change classification. A disruption might start as light, then as time passes and the alleviation is difficult or impossible the disruption evolves to mediocre. If the disruption is still unresolved the now mediocre disruption may become severe and possibly total. On the other hand a disruption first classified as mediocre may be reduced to light if local or immediate alleviation is effective.

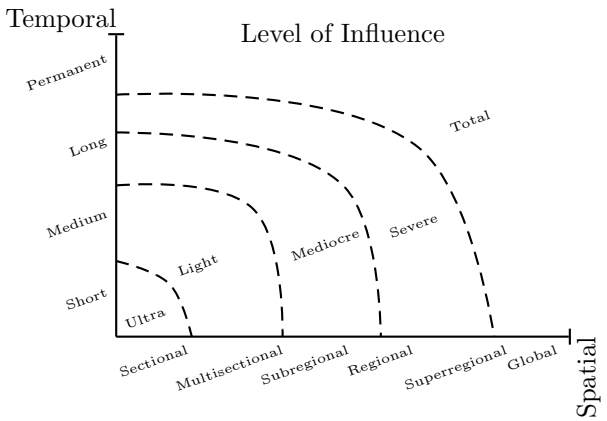
2.2.3.3 Comparison

Taking a closer look at the two approaches to a classification scheme it becomes clear that spatial and temporal locality is based on the "what" and the "when" of a disruption. On the other hand the LoI concerns the "how" of a disruption.

In a slightly different formulation the spatial locality is concerned with what is part of the disruption, the temporal locality distinguishes when the system is disrupted and LoI tries to indicate how the disruption should be alleviated.

This supplementary symbiosis can be split into two indicating characteristics and an alleviation characteristic – the localities and LoI respectively. The indicating characteristics are most likely immediately known where as the alleviation characteristic is based on a possibly automated evaluation of the current status of the system.

As a small example consider a section where a car breaks down reducing the capacity to 40% of normal. If the current load of the section is below the new



The dashed lines can be considered as iso-LoI lines

Figure 2.3: Classification guide

impeded capacity the LoI is ultra-light. If the load is above 40% vehicle dispersion is necessary. The influence then becomes light if the dispersion is locally possible – there are other local sections that can be used. If local dispersion is impossible due to high load on the other sections the LoI increases. In extreme cases, where high capacity sections are severely impeded, severe or total disruptions may occur when the flow capabilities into an area are significantly less than vehicles destined for or passing through that area.

The classification is divided into three parts, which can be visualized. Figure 2.3 is a depiction of the classification terminology and the general coherency. The coherency is only a guideline and some cases will fall outside it. As an example the permanent closure of a section of no importance to the general flow through the surrounding area is at most light and not severe or total as indicated by the guideline.

2.3 Scenarios

Having the classification scheme in place allows us to discuss the different scenarios and classify the different disruptions. The scenarios below have been chosen because they exemplify the function and impact of ABIT in real world situations. Each scenario describes a situation and what we expect the ABIT system to be capable or aware of. The scenarios have been divided into sections based on their cause. The most commonly accepted disruption – an accident – is discussed first. This is followed by more detailed scenarios with emergency

vehicles and imposed, planned, external and secondary disruptions.

2.3.1 Accidents

A traffic accident usually reduces the capacity of the affected area. This means that the spatial locality of an accident is small – sectional or in large accidents multisectional. In light traffic the queuing delay is negligible, but under heavy traffic the capacity of the area may be reduced below the amount of incoming traffic causing theoretically infinite queuing delays.

The intent of ABIT is to alleviate the queuing by diverting the traffic from the accident. Today this is done by the radio simply proclaiming that the accident has happened and that it will be some time before it is cleaned up. It is then left to the individual driver to react to the information.

Assume the simple transport network depicted in figure 2.4. In intersection x_2 an accident has occurred severely reducing the capacity. The vehicles directly inbound on x_2 cannot be diverted, but all vehicles before intersection x_1 can. As soon as the reduction in capacity is acknowledged by the system the vehicles before x_1 are assigned new routes. These routes are based on the new capacity situation and will thus allow some to continue on their original route. In the example vehicles destined for areas A and C are preferably diverted and vehicles heading for area B are preferably sent through x_2.

From the systems point of view this leaves many unanswered questions. Among these are how to:

Detect the disruption. As the system is aware of every vehicle through the agents different detection schemes can be used.

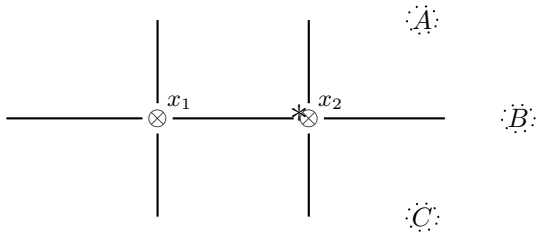


Figure 2.4: Simple transport network

- Manual detection based on reports by the police or other authorities
- Automatic detection by the system based on the monitored throughput of an area
- Automatic detection by monitoring reduced traveling speed by individual vehicles

Of the above three the last two are preferable to use in the case of accidents as the system automatically detects when the capacities are reduced or restored. The problem is that it might require complex algorithms to decide what is happening and how long it lasts. The Canadian Transport council among others has for a while been addressing this through cellular tracking [Hill and De Santis, 2002].

When a capacity reduction is considered, a disruption is dependent on detection and model. Under normal traffic the flow of a section may fluctuate to some extent. To avoid generating unnecessary disruptions this must be considered when designing the interaction between detection system and guiding system.

It might be possible to apply forecasting models from traditional traffic research to predict possible disruptions making the system enter a proactive role instead of reactive.

Find the new route. To find the new route a general routing problem has to be solved. Before the disruption some guiding scheme was in effect aiding in the route selection. The accident has imposed new criteria on the scheme and it has to be recalculated. When the recalculation finishes the new scheme is used to select paths for all vehicles in the system. The interaction between the scheme and the path selection, the dispatch plan, is complex and will be covered separately in chapters 4 on theory and appendix B on the more practical aspects. While the new scheme is being calculated, some immediate dispersion strategy has to be applied.

Ensure diversion of traffic. As the drivers are individuals there is no obvious way to accomplish this. The quality of the routing should in time teach the drivers to follow the guiding. If this does not suffice, route pricing could impose extra cost to drivers deviating from their assigned path, if it introduces excessive negative impact on other commuters.

In the following we will cover four specific accident scenarios. Each of these will in detail discuss the expected actions of the system.

2.3.1.1 Fast recovery

Two vehicles collide, but are still able to move. The drivers move their vehicles to non-impeding positions within short time thus exhibiting short or medium temporal locality.

From the driver's point of view the accident happens and they both move their vehicles to the non-impeding positions. When they are ready, they reenter the system as their route is kept updated.

In the case of fast recovery it might not even be considered a disruption. If the impact is so little that the detection system does not indicate a disruption then nothing more happens. The system never reacts on it.

If the detection system on the other hand indicates a disruption then the guiding system immediately assesses the situation. If the capacity reduction is relatively significant immediate dispersion is inaugurated and the guiding system begins to generate the new routing scheme. While the new scheme is being generated the old is used except for the vehicles immediately inbound on the affected section. The gap between new and old routing schemes should be reduced by using incremental algorithms thus avoiding unnecessary suboptimal usage of immediate dispersion.

If the capacity reduction is small relative to the current load of a section no immediate action is taken, only the general routing scheme might be updated.

When the capacity is recovered the routing scheme is regenerated and the immediate dispersion is stopped.

2.3.1.2 Slow recovery

In this case the capacity reduction is not immediately recoverable as the temporal locality is longer – medium to permanent. Vehicles may have broken down or the infrastructure may be damaged. Automatic detection can not detect the duration, which could be important when alleviating the disruption.

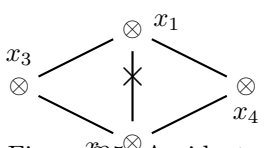


Figure 2.5: Accidents

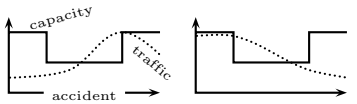
When the accident happens the vehicle logically leaves the system and the new capacity of the

section is assessed and used to generate a new routing scheme.

Given the network in figure 2.5 on the preceding page an accident happens on $\overrightarrow{x_1x_2}$. If the new capacity $\lceil\overrightarrow{x_1x_2}\rceil$ is reduced and the load $|\overrightarrow{x_1x_2}|$ is less: $|\overrightarrow{x_1x_2}| \leq \lceil\overrightarrow{x_1x_2}\rceil$ the disruption is assumed ultra light or in the case of automatic detection simply undetected as no queuing occurs. If on the other hand $|\overrightarrow{x_1x_2}| > \lceil\overrightarrow{x_1x_2}\rceil$ the disruption will be detected as traversal speed will be reduced due to queuing. Dispersion is necessary and if it can be done sufficiently via x_3 or x_4 the disruption is classified as light. If local dispersion is insufficient due to inadequate combined local capacity from x_1 to x_2: $\lceil P(x_1, x_2)_{local}\rceil$, the LoI is increased according to the size of the area used for the diversion. The diversion area intrinsically defines the LoI as indicated in 2.2.3.2 on page 16.

So far we have assumed static inbound flow while the disruption is alleviated. While this may be true in many cases, dynamic inbound flow has to be considered as well.

Figure 2.6 shows two situations where the inbound traffic changes during the disruption. In case a) the disruption is at first ultra-light, but as the inbound flow increases it becomes more and more significant. Here it would be preferable if the system could reduce the period of congestion by diverting individual vehicles destined for the disruption sufficiently early. This would also reduce the load on the immediate surrounding area.



a) Increasing b) Decreasing

Figure 2.6: Inbound traffic

In effect any detected disruption would immediately start diverting traffic at any distance from the destination if their route was going through or close to the effected area. Even though this would be a good idea to ensure smooth traffic it might overcompensate and thus both under utilize the infrastructure and send vehicles on suboptimal routes.

Overcompensation is exactly the problem in case b). When the accident happens the impact is immediate. The capacity is reduced significantly below the incoming traffic and thus queuing occurs. This is handled by local remedies and dispersion in larger areas is commenced. The important issue in the diversion is that the incoming traffic will drop below the reduced capacity allowing normal operation. With global diversion we might divert vehicles which will not be affected by the accident because when they arrive at the scene the incoming load is below the reduced capacity.

The complexity is definitely not simplified by the fact that we have no idea

exactly when accidents happen or when they are recovered. This will be a key issue to making ABIT an acceptable system. Precise estimation of the new capacity and prediction of load over time is essential to address this issue properly. Traditional models may be applied to enhance the system by forecasting and pre calculation. Post incident analysis may be used to establish a forecasting scheme or accident avoiding measures.

2.3.2 Emergency vehicles

These incidents are short, roaming and sectional as they usually only concern one vehicle at a time. Nevertheless, if they interfere significantly with the capacity of an area they are considered a disruption.

The sectional roaming of an emergency vehicle might make them difficult to incorporate into the global routing scheme. The agents can instead be used to advise the drivers on incoming emergency vehicles. The system may ask other drivers to hold right or left and estimate the time until the emergency vehicle passes. The emergency vehicles might very well be headed for an incident that is already considered a disruption in the system.

It would be a possibility to use ABIT to ensure swift access to the site. The vehicle enters the system as any normal vehicle with a designated destination. The major difference is that it is suggested a path based on greater flexibility in the network.

2.3.3 Imposed disruptions

So far we have looked at strictly internal disruptions. Disruptions that are caused by special or malfunctioning vehicles. This is certainly not the only type of disruptions that has to be considered.

The next type considered is imposed disruptions. These are capacity reducing events that are not advance knowledge and not necessarily caused by vehicles. Examples of such are blockades and demonstrations, which can have heavy impact on the capacity of the affected area.

Whenever the disruption occurs the situation is assessed. Dependent on the immediate classification of the disruption proper action is taken.

From the drivers perspective all that happens is that a new path is proposed.

This can be accompanied by more elaborating information on the reason for the route change.

The system however has far more work to do. The problem for the system is to find the best routing of all vehicles based on the assumed classification of the disruption. This is done by updating the general routing scheme along with immediate dispersion in the local area. If the disruption is not roaming and not periodical the system has reached a steady state and until normal capacities are restored nothing happens. On the other hand, if the disruption is both roaming and periodical, extreme care has to be taken to achieve the best routing. The pitfall is to reroute distant vehicles from the currently affected area into areas that are affected when they arrive. This might make an inconsiderate system perform worse traffic routing than unguided traffic.

2.3.4 Planned events

The following type of disruption is not much different from an imposed disruption. The key point here is that imposed disruptions are not advance knowledge and their classifications are hard to estimate. Planned events are classified to some extent before they occur and the only reason that they still are considered disruptions is that their impact is temporary.

Upon the identification of a planned event by the system the new routing scheme for the disruption is precomputed. It is not put in effect immediately, but should be introduced into the system in such a way that only vehicles affected by the disruption will adhere to it. The crucial points here are the possible time window of the event and the propagation speed of vehicles through the network. These have to be carefully combined to ensure good routing.

Planned events cover infrastructure based sport events, like the Monaco Grand prix and the New York Marathon, planned demonstrations and the like. Shorter construction work where the capacity reduction can be considered as temporary relative to the routing scheme will fit into this category as well.

If the duration of the capacity change is long, it might be considered an actual change in the infrastructure instead of a disruption.

2.3.5 External disruptions

The above disruptions can be described as internal to the system. They directly affect some specific part of the system and are generally controlled or inflicted by humans. The attention is now turned to external or general disruptions.

These are any capacity reducing event caused by less controllable sources: heavy rain, snow, ice, earthquakes, hurricanes, etc. The classification can be any of the above combinations and the guideline is effectively useless here.

These disruptions are identified separately as these are most likely the only disruptions that can be roaming and super regional or larger at the same time.

This type of disruption is considered the hardest to recover as they can exhibit all the most complex interferences with both infrastructure and vehicles.

One approach to this is to use automated capacity detection by monitoring vehicle traversal speed through out the network. This way the immediate condition of every section could be considered. As with imposed disruptions great care has to be taken to avoid a negative impact of using ABIT.

2.3.6 Secondary disruptions

To close the section on different disruptions we add a final type of disruptions that we call secondary disruptions. These are special by the fact that they have no evident and direct cause, but still reduces the capacity of an area. The key issue is that no single event has lead to the disruption. If a disruption is caused by any single other disruption it would not be considered a secondary disruption, but included in the causing disruption by altering the temporal or spatial locality of it.

Several different incidents like a planned lane closure, an emergency vehicle and weather conditions may in combination give a secondary disruption.

A typical secondary disruption is a traffic jam. Another example can be the stop-and-go waves on congested highways.

Interestingly these waves were predicted and solved several decades ago in [Lighthill and Whitham, 1955]. The solution was also stated in the article: “Increase the headway”.

Secondary disruptions can have numerous reasons for occurring. Some of them are caused by a combination of poor routing and insufficient areal capacity along with inconsiderate driving. This again may be indirectly generated by a combination of other disruptions. The point is that a secondary disruption is a disruption that has no distinct reason. A disruption caused by a knock on effect of a single other disruption simply increases the LoI it does not constitute a secondary disruption.

A goal of the system is to avoid these secondary disruptions as many of them are caused by insufficiency in the utilization or construction of the network.

Some of them are insolvable. Consider $2n$ vehicles wishing to park in a parking lot with only n spaces or a roundabout. We will discuss the roundabouts later in this text. Hopefully the system might help traffic forecasters to alleviate this kind of problems before they become critical.

2.4 Complex traffic systems

To this point only the direct guidance or route planning of the drivers has been considered. An important possibility to consider is the usage of existing infrastructural control together with ABIT.

2.4.1 Enforced control

Today all traffic systems have enforced control of some kind. This is usually simply stated as traffic law; do not pass red, do not go the wrong way on one way streets and such.

These rules are necessary for the traffic to be safe and effective to some degree. What is interesting is to identify which of these laws that can be used by ABIT. Today this is already in effect. Tunnels and highways are made with rush hour lanes that can change direction based on the direction of the traffic at different times of the day. Light controlled junctions are programmed to make "green waves" that increase the capacity of an area in a specific direction. In southern Spain an entire road (50km) is manually altered by the authorities to accommodate the extreme weekend and holiday traffic from Granada to the coast.

Today these measures are effective and gives much better service to the drivers.

The extension with ABIT will be to let the system decide when rush hour lanes are necessary and how the light controlled junctions are to be programmed. The question to answer here is whether ABIT can be used to provide even better service based on the same infrastructure.

Major cities in the USA are currently using dynamic programming of enforced control to get better utilization and service based on the same infrastructure.

2.4.2 Subtle control

Besides from enforced control, which is secured by the authorities, ABIT can be used to increase the awareness of the driver. Information like average speed on the current section and the next section, road conditions, upcoming queues and distance to end of these might all make the traffic safer. Even detailed information like "non braking speed to pass next green", which is used some places in Germany, can possibly increase the capacity of a section.

The idea here is to use the increased monitoring and data gathering beyond the routing and present it properly to the driver.

2.4.3 No control

Sections with no control are sections that exhibit some complexity in the infrastructure while there is no enforced control. An example could be a roundabout.

The application of ABIT to a roundabout is highly dependent on the actual layout of the roundabout. If it contains several lanes the ABIT can guide the driver to select the right lane to minimize intra-roundabout lane changes. It can also ask, but not enforce, drivers to wait for a while before entering if exiting the roundabout might cause problems for other vehicles.

As the section title states this is uncontrollable and non enforceable, but it might still be of help to the general traffic.

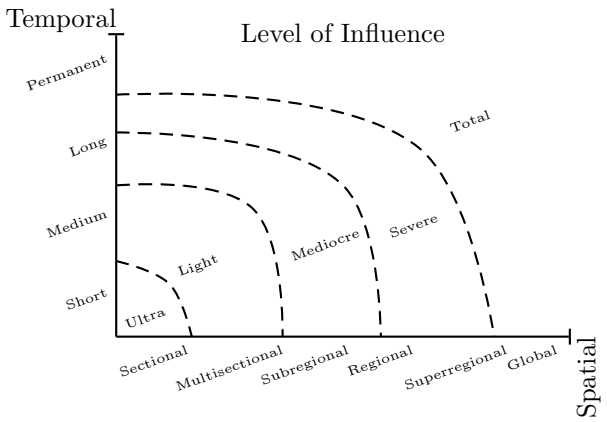


Figure 2.7: Classification guide

2.5 System

Several types of disruptions were covered in the preceeding discussion and it has become clear that the system has to act almost identically to all disruptions.

1. Detect the disruption
2. Commence generation of new routing scheme
3. Inaugurate immediate dispersion strategies
4. Upon completion of the new routing scheme it has to be propagated properly into the network
5. When the disruption is alleviated or the classification changes repeat step 2 to 4.

2.6 Summary

In this chapter we have discussed the definition and classification of a disruption in the ABIT system. The three characteristics of a disruption can be combined into a guideline given in figure 2.7.

Several disruptions were considered and besides normal operation the following terms were defined:

User imposed deviation caused by the autonomous nature of the driver,

Accidents of any kind,

Imposed disruptions such as demonstrations or other deliberate human interference with the infrastructure,

Planned events like the New York marathon or the Monaco Grandprix,

External disruptions caused by the impact of nature, and

Secondary disruptions resulting from the mere presence of to many vehicles.

Furthermore it was briefly covered how the increased flow awareness through ABIT or Pervasive Traffic Intelligence (PTI) could be used in complex traffic systems to increase the efficiency of these.

The next chapter on history will contain a more detailed history on the traffic assignment in Operations Research (OR) and Traffic Science (TS).

Chapter 3

History

This chapter contains an overview of the evolution of Intelligent Traffic Systems (ITS) within Traffic Science (TS). The theoretical milestones will be covered and the latter part of the chapter will describe the evolution of the major technologies of practical TS.

3.1 Theoretical evolution

TS emerged into a separte field a few years later than OR. Through the first years tractability was an issue as the analytical properties of the problems required large-scale computations by the standards of that time. As the computers became faster and more able, the research grew significantly and flourished in the 1980'ies, where state-of-the-art research was able to actually aid decision makers within reasonable time.

The field operates with demand and supply. Most problems are considered as producer/consumer problems. The producer is the infrastructure which generates a supply which is utilized by the users or customers to fulfill their demands of travel. For any model to exist the populace is considered as homo economicus – the rational human. This implies that every user chooses his or hers most

suited alternative fulfilling the users requirements. In discrete choice modeling utility is a measure of how well suited an alternative is for the given requirement. Homo economicus will thus choose the alternative with highest utility.

If user A requires to travel from A_o to A_d for some reason, user A represents a transportation demand from A_o to A_d. User A considered as homo economicus will then choose the supply alternative – or path in our case – that represents the best utility. Utility is used as a generalized term instead of just shortest, fastest or easiest as individual users might have different perception of utility.

A simple utility-function consists of different constituents or variables, that are related to the alternatives and taste parameters related to the users.

$$U_{n,m} = \sum_i \Theta_{n,i} x_{m,i}$$

Here, $\Theta_{n,i}$ is the taste parameter for individual n variable i and $x_{m,i}$ is the value of variable i for altertive m. As both the $U_{n,m}$ and the $\Theta_{n,i}$ are not known the utility function is calibrated by estimating the values of the $\Theta_{n,i}$. This requires a finite set of alternatives and a sample of individuals having decided between the different alternatives. A maximum likelihood calibration is then applied to find the set of $\Theta_{n,i}$ which is most likely to result in the sample. This, however, is a complicated task and within estimation and forecasting much research is concerned with properly modeling the taste parameters.

The effort required to design and calibrate the models is essential. The models have to be comprehensive in the sense that every significant variable is in the model. The models usually contain high complexity as the variables must be represented correctly.

If a proper statistical model is constructed it can be used for a wide variety of issues. Most importantly for us, it can be used to forecast the situation in the case of limited changes. Hence, we can use a properly constructed and calibrated statistical model to forecast the expected real time status of an area in advance.

The construction, calibration and forecasting of statistical models, however, is not the primary topic in this thesis. [Ortúzar and Willumsen, 2001] covers these topics in further detail for the interested reader.

The original Traffic Assignment Problem (TAP) is constructed to find the distribution of traffic in the infrastructure. The solution to the problem is the distribution of the demand through the network assuming that the demand is fixed. Furthermore, the distribution found must satisfy Wardrop's first principle.

Minimize	$\sum_{\forall sections}\left[\int_0^{\lvert\overrightarrow{ab}\rvert} f_{\overrightarrow{ab}}(x)dx\right]$	Nash equilibrium
	$\lvert P(x_a, x_b)\rvert > D_{a,b} \quad \forall a, b$	Demand satisfaction
	$\lvert p(.)\rvert \geq 0 \quad \forall$paths	Non-negative path flows

Above, $D_{a,b}$ represents the traffic demand from node x_a to node x_b. The integral is introduced by the Beckmann transformation [Beckmann et al., 1956], which ensures that the problem solution is actually the Nash equilibrium.

The static requirement of demand in TAP is very restrictive and many real world instances do not have this. An example is the modelling of rush hour where the demand increases and then decreases over a period of time. Simplifying this to a constant demand cannot properly represent the specific demand distribution. [Yagar, 1971] introduces a changing demand in the considered time period and the corresponding problem is called Dynamic Traffic Assignment (DTA).

Even though the DTA yields far more representative solutions to the rush hour problem evolvement in ITS has shown a demand for even more representative models and methods. The Consistent Anticipatory Route Guidance (CARG), mentioned [Dong et al., 2006], is one such application specific problem definition. Essentially the CARG is a modified DTA. The major difference is that information from the network is accumulated and applied during the solution. This might seem confusing as the intent usually is to solve the problem before or without actual realtime information. The motivation for CARG is that it now is possible either by simulation or actual real time status information to incorporate the state of the infrastructure dynamically into the solution process. Obviously this is very complicated and consequently CARG instances are difficult to solve.

The TAP, DTA, and to some extent CARG have all been formulated in many different ways. Due to this multitude of different formulations addressed within TS, [Psaraftis, 1995] proposed a classification approach to TS problems.

The scheme in part 4 of [Psaraftis, 1995] classifies a problem based on 14 different characteristics. The main idea is that if you have two different problems which are classified alike, the algorithms usable for the first problem is usually also usable for the other problem. On the other hand if the two problems differ in just one of the characteristics it is unlikely that the same algorithm is usable for both.

The 14 characteristics are:

Information type: Is the information dynamic or static? If the information is static the solution for a problem is well founded. The information could also be dynamic thus requiring more elaborate understanding on both solver and model. In this thesis, the information is dynamic.

Time: Is representation of time within the problem important? When optimizing container packing it is usually not taken into account how long it takes to pack the container. On the other hand when considering traversal in physical networks the time is essential. At some point in time a section is filled with some vehicles which at a later time is on another section. This is the case with Agent Based Individual Traffic guidance (ABIT) and the time dimension is thus essential here.

End: Has the problem a definable end? Any problem with a definable end has a specified end state. Once this is attained the given solution to the problem can be evaluated. However if no such end is definable the concept of a solution is significantly different. ABIT as application has no specific end as it is meant to be continously online.

Information Quality: Is it vital to have perfect information or is some noise acceptable? Routing trains tightly through an intersection allows absolutely no noise as the trains will collide. However pinpointing the position of a single vehicle in a fleet down to a square centimeter is in most cases unnecessary. In ABIT high quality of information is desirable. Due to the autonomous nature of private vehicular traffic it is unlikely that actually perfect information will be usable for anything else than calibration.

Event priority: Are some events more important to the problem than others? Prioritization of events can be important if considering livestock transportation. In our case all events are treated equally.

Information updates: How is the information gathered and updated? Here the information is intrinsically communicated and updated in real time. In other cases the information could be supplied as advance knowledge or at specific checkpoints in the problem definition.

Information availability: Where, when and how can the information be accessed? The Central Information System (CIS) automatically ensures that all vehicles are equally able to access information regarding the infrastrucure. The information is thus almost always attainable.

Path assignment: How can the solution be modified? Path assignment defines how a new solution can be generated. In some cases paths and vehicles can simply be removed or inserted to find a new solution. In ABIT all vehicles are physical entities and it is not possible to directly remove and reinsert them later.

Solution time: How fast is the solution needed? This considers the runtime limitations for finding a solution. Here we must be able to redirect immediately inbound traffic as quickly as possible. The faster we achieve good dispersion the less impact the disruption will have. In the presence of a disruption the traversal speed of any vehicle is reduced and it may be possible to allow a few minutes to find good solution. On a high capacity section loaded with 6.000 vehicles per hour, 2 minutes delay potentially sends 200 vehicles in wrong directions. This should at all costs be avoided and the solution time acceptable must be evaluated, possibly automatically, at every disruption.

Postponement: Can a demand be postponed? In our case the demand is an actual vehicle on the roads and postponement is thus impossible.

Dynamic objective function: How is dynamism in the problem represented? For ABIT different models can be applied. If we choose a time expanded classic TAP the objective function is only extended by extra variables. The objective function is thus not significantly changed by the inclusion of dynamism. Otherwise we can choose a integral representation of the objective function which directly represents the dynamism through integrals over time.

Time limits: Some problems have specified deadlines or time limits on some parts of the solution. Within the ABIT system such a demand could be devastating to the general routing as a few specific trips would require dispersion of many others. Thus in ABIT we accept no strict individual trip time limits.

Fleet size: How many vehicles are considered? In ABIT we consider all ABIT-enabled vehicles, which hopefully in time will be all. The considered fleet size is thus huge.

Queuing: Is it necessary to represent queuing properly? Considering only low load traffic problems causes only negligible queues. On the other hand the importance and secondary effect of queuing is crucial once the load approaches congestion. Thus in ABIT it is evident that queues must be represented realistically.

The characteristics summarized in figure 3.1 on the next page are essential in the decision of which model and solver to use.

The crucial part in any model independent of which tradition it originates from is the parts which actually represent the real problem. These parts can be divided into two:

Characteristic	ABIT
Information type	dynamic
Time	essential
End	not implicitly defined
Information quality	high
Event Priority	equal
Information updates	intrinsic
Information availability	global
Path assignment	reassignment only
Solution time	limited
Postponement	not possible
Objective function	non dynamic
Time limits	none
Fleet size	huge
Queuing	realistic

Figure 3.1: Classification

Constraints which ensure that solutions accepted by the model are actually feasible solutions to the real problem

Objective function or cost function which represents the benefit or loss from a given set of values for the model variables.

Basically, both are defined depending on the choice of variables. Great care has to be taken to ensure that the constraints are representable with a given choice of variables without resulting in intractability of the model. Choosing variables that yield simple constraints might result in overly complicated objective functions – and vice versa.

At this point however, and for traffic modelling in general, it is far more important how the cost function is defined.

3.1.1 Cost function

Generally the cost function used for traffic assignment in TAP or DTA is simpler than the equivalent for statistical modelling. This is due to the numerous computations involved and the rigid mathematical requirements in traditional solvers for traffic assignment.

Usually a full scale statistical model is constructed and the most significant

variables extracted. [Bonsall and Parry, 1990] as well as [Pang et al., 1999] are among several papers that indicate that the most significant variables in the percieved cost af a given path are the travel time and travel distance.

This has spawned a wide range of research aiming at finding the statistically best model to represent the relationship between flow and travel time. This research is not solely driven by the statistical significance of time, but also by the computational possibilites of computers and solution methods.

Examples of generic cost functions are:

[Smock, 1962] focus on the traversal time of a section for a Detroit study:

$$t_{|\overrightarrow{od}|} = fft_{\overrightarrow{od}} \exp^{|\overrightarrow{od}|/\left\lceil\overrightarrow{od}\right\rceil}$$

where $fft_{\overrightarrow{od}}$ is the free flow traversal time of the section $\overrightarrow{od}$.

This function is a simple smooth approximation that exhibits exponential growth in traversal time. The cost function has in this version only one calibration parameter and thus calibration is simply, but coarse.

[Department of Transport, 1985] proposes the following general form. It is intended to increase the precision of calibrated instances:

$$t(|\overrightarrow{od}|) = \begin{cases} d/fft & |\overrightarrow{od}| < F_1 \\ \frac{d}{fft + \frac{fft-cft}{F_1-F_2} - \frac{fft-cft}{F_1-F_2}|\overrightarrow{od}|} & F_1 \leq |\overrightarrow{od}| \leq F_2 \\ d/cft + \left(|\overrightarrow{od}|/F_2 - 1\right)/8 & |\overrightarrow{od}| > F_2 \end{cases}$$

Here F_1 is the maximum load for the given section where the free flow time fft is still valid, and F_2 is the capacity where the stable capacitated flow time cft is valid.

This stepwise smooth and connected function is capable of considering representing three phases for the section a) light load, b) onset of congestion, c) congestion. This obviously represents reality differently and most likely better than above, but the function is computationally more cumbersome to use.

[Bureau of Public Roads, 1964]

$$t(|\overrightarrow{od}|) = fft\left[\left(1 + \alpha\frac{|\overrightarrow{od}|}{\left\lceil\overrightarrow{od}\right\rceil}\right)^{\beta}\right] \tag{3.1}$$

Here, fft is as above and α and β are parameters used for calibration.

The addition of the calibration parameter on the exponent allows for better differentiation between section types as the impact of increased traffic is dependent on the actual section and intersection layout.

[**Meschini et al.,**] gives the following model of the perceived cost:

$$w_{o\to d}(\tau) = \Theta \ln \left(\sum_{r \in R} exp \left(\frac{-c_r(\tau) + w_{HD(r)\to d}(t_r(\tau))}{\Theta} \right) \right)$$

Here, R is the Efficient Forward Star (EFS) from o to d. The EFS is the set of directed sections bringing a vehicle closer to it's destination. The Θ is a scaling parameter and the numerator is a utility measure for free flow and interfering flow on a section. This can be used to indicate attractive routes when trying to divert traffic.

Of the above different cost functions function 3.1 on the preceding page has been generally accepted as the best approach.

Each of the above cost functions have different qualities. Some of these represent low flow situation very well, while others are specialized at congested situations. If we consider the possible solution methods the crucial difference in the cost functions, however, is whether they are separable or not.

The term separable is used if the cost function exhibits a special mathematical property. This is when cost of traversing a section according to the cost function is indifferent to the flow on any other section or the opposite direction on the same section. The cost of traversing a section in one direction is thus solely given by the flow on the same section in the same direction.

This is a very attractive mathematical property and comparing the solution times for separable cost functions with inseparable cost functions clearly shows the importance of separability. A separable cost function almost automatically leads to usable results. On the other hand an inseparable cost function can result in unsolvable instances. [Jara-Diaz and Friesz, 1982] points out that a line integral does not have an unique, unambiguous value unless the Jacobian matrix formed from its integrand is symmetric. These symmetry requirements are unlikely in our real world settings. This is especially true for inseparable cost functions.

The drawback of a separable cost function however is significant. Allready in [Prager, 1954] it was made clear the inseparable cost functions would be required. For many scenarios applying a separable cost function will be ridiculous. A simple example would be the interaction at a simple uncontrolled T-junction.

Dependent of the traffic laws in effect some traffic will be required to yield for a flow on another section or another direction. For low load this might not be significant, but as the traffic pressure increases the separability will fail to estimate the yielding delay properly. The yield delay is calculated based on flow that is not on the same section as the yielding flow and a representative cost function will thus be inseparable.

Calibrating a perceived cost function is a tedious task. It requires careful audit and survey of commuters in the case of a known change to the network. At present the level of detail attainable is inadequate to sufficiently calibrate complicated cost functions. However once ABIT or any Pervasive Traffic Intelligence (PTI) is active the information necessary for calibration can be accumulated and refined automatically.

3.1.2 Perspectives

As the industry has begun applying the research three areas of TR applications have been commonly identified.

The strategic perspective is the largest scope. It considers long term alterations. Typical decisions in this area is whether to build new roads or simply increase the capacity of existing sections. As building 1 km of motorway is extremely expensive the investments involved are staggering. The Copenhagen expansion of high capacity beltway is estimated at 2 billion DKR (> 300 million USD). Extensive strategic modeling exemplified in [Rich et al., 2003] has been used to forecast traffic demand and flow distribution to ratify the expense.

Narrowing the scope to the next area, called the tactical perspective, we consider which changes we can do to existing road infrastructure to increase the yield or resulting supply. This includes changing green-periods or green-waves in signalized areas. Other applications includes the addition of designated turning lanes or special right-of-way rules. The line between the strategical and the tactical is fuzzy, but a rough approximation is that the longer it takes to implement, the more likely it is a strategic problem.

The third area, the operational perspective, is distinctively different. This considers real time problems and real time information and solutions for immediate application to the traffic.

Considering a town with no prior traffic planning, practitioners should first conduct a strategic analysis of the infrastructure. From this they might conclude that no significant infrastructure alterations are necessary or viable. The next

would then be a tactical study on optimizing the supply of the infrastructure. This could lead to changing the green light periods or changing lanes layout.

Finally an operational perspective should be attained for the everyday optimization of the flow distribution.

All of these perspectives affect the supply of the infrastructure and thus the perception of routes to the drivers. Most of these we do not notice. Both the strategic and tactical perspectives are efficiently hidden in government reports and high priced consultant work. The drivers usually simply perceives a supply increase over some time span.

The operational perspective, however, has been far more visible to the drivers for decades. A simple example is the traffic radio. It efficiently informs the travellers of the real time information of the infrastructure thus allowing the traffic to redistribute given the new conditions. Traffic radio however has the inherent problem that it is not possible nor advisable to include all information. Thus only the most significant information for the entire commuting populace is communicated, not the actually important information for the individual driver.

Another example is in the case of significant network alterations or dangerous situations where usually the police is manually controlling the traffic. Recent years has seen the development of effective and necessary applications as covered in section 3.4 on page 50 on applied practice.

3.1.3 Models

It is evident that the processing and information requirements are very important. It is of no use to have a perfect result, if it takes two hours to get it. The time window we are considering is significantly smaller. Vehicles immediately inbound on a disrupted section require a new route within a few seconds. For vehicles traversing routes that are influenced by the disruption the window may be larger. It is essential that we get a good solution fast and that we are able to improve on it given more time.

It is also important to realize that we at any point in time have a running solution. Vehicles are traversing the network are thus implicitly displaying a feasible solution. Our goal is not necessarily to generate a new solution from scratch. It could also be to isolate the implicit solution and get from it to the near optimal or optimal solution.

We need a solver that can find Wardrop's equilibrium or Wardrop's social equi-

librium [Wardrop, 1952] for a given network state. This requires that Wardrop's equilibria can be expressed in the model.

The crux of Wardrop's first equilibrium is that all used paths from an origin to a destination has the same perceived cost. The social equilibrium, Wardrop's second equilibrium, requires the total perceived cost for all commuters are minimized. In most traffic networks today these equilibria do not coincide.

Travel distance is static but time, as indicated in appendix C is a much more difficult measure to utilize. The problem is that the time it takes to traverse a path is dependent on the number of vehicles on that path. Even worse, vehicles on other paths influence as well.

More explicitly the travel time of a path is defined by the travel time on each section and intersection on the path. Every other path using the same sections or intersections will influence as well. The complexity is certainly not reduced when the relationship between flow and perceived cost is defined by a polynomial function as in equation 3.1 on page 35.

In the perfect theoretical world every driver will do as we ask, but in the real world the autonomous nature of vehicular transport is troublesome to model. To alleviate this we might have to generate robust solutions that can assimilate minor behavioral differences without offering a significantly reduced service on some paths.

Assume that we have two possible routes. One route, A, containing small roads and another B utilizing only trunk roads. The optimal distribution will send 100 vehicles on A and 2000 vehicles on B. This solution however is not stable. The 100 vehicles utilizes A very close to congestion. Hence, every excess vehicle will severely deteriorate the traveling speed. Only 5 vehicles more will lead to severe congestion due to oversaturation and thus a poor solution for the vehicles that chose route A. The trunk route B however is far less sensitive to excess traffic. Even though congestion will appear here too, the effect of congestion is less per vehicle. A more robust solution to distributing the vehicles will thus be sending only 90 or 95, dependent on the required degree of robustness, on route A and the rest on route B. This will allow more vehicles to select route A despite contrary routing information without leading to complete congestion on the route.

When considering the approaches we have to find a better solution within few seconds and a near optimal solution in a few minutes. Given the required robustness of the solution and the uncertainty in even short term prediction of traffic it may be of no relevance to actually find the optimal solution at all. For theoretical purposes, however, it will be interesting to find the optimal solution.

The model for perceived cost above indicates that the modeling field we choose must be flexible enough to handle the mathematical intricacies as well as allow fast and good feasible solutions.

For the convenience and safety of the drivers we cannot continously or even daily change the assigned route.

From the tradition of mathematical modeling a wide selection of modeling methods exist. The following section contains a brief overview of the methods primarily applied.

3.1.3.1 Linear Programming

Linear Programming (LP) is the model type which had and still has the greatest impact within Operations Research (OR). It is by far the model type with the strongest solution methods. A wide variety of both commercial and open source solvers are available today.

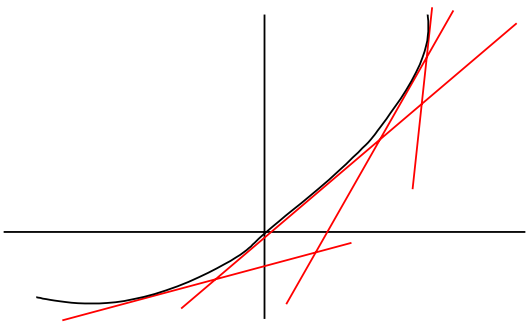
Every mathematical program consist of one or more objectives and a set of constraints. For the remainder of this thesis we will consider only single objective problems. The term linear refers to that the objective function and all constraints must be linear in the variables.

A small example is:

Minimize	$2x + 2y - 4z$	objective
Subject to	$x + 2y + z \leq 8$	first constraint
	$-x \quad + z \leq 0$	second constraint
Where	$x, y, z \in \mathbb{R}_0$	variable definition

Modeling capacity is restricted as many real life constraints are non-linear. On the other hand most realistic problems that can be expressed in LP are quickly solved.

A method to approximate a non-linear model as LP is by constraint expansion as depicted in figure 3.2 on the next page. The idea is to aggregate a convex constraint with several linear constraints. This makes is possible to approximate a nonlinear constraint with linear constraints and thus not violate the LP requirement. The feasible space defined by the linear constraints can then be defined arbitrarily close to the original constraint. This is done by increasing the number of substituting linear constraints.



The figure shows how the black non-linear constraint is approximated by a set of red linear constraints.

Figure 3.2: Constraint expansion example

The evident problem here is that the number of constraints increases rapidly with both system complexity and precision in the model.

LP is insufficient when the linearity cannot be justified or the variables are not continuous. Furthermore all variables may take non-integer values. The result of this is that a solution might require 9.56 vehicles to follow a specific path. Even though attractive we are not allowed divide a vehicle.

However it may prove a very efficient tool for solving subproblems in a larger solution complex.

3.1.3.2 Integer Programming

The "9.56 vehicles" problem is solved using Integer Programming (IP) or Mixed Integer Programming (MIP) where respectively all or some of the variables are required to attain integer values. Thus solutions to IP problems will not attempt to divide individual vehicles.

The introduction of integer constrained variables also increases the modeling strength significantly as strict conditional and logical constraints can be introduced. The constraint $x \geq 0 \Rightarrow y \leq 0$ can for instance be expressed as:

	$x - dM \leq 0$	$x \geq 0 \Rightarrow d = 1$
	$y - M + dM \leq 0$	$d = 1 \Rightarrow y \leq 0$
Here	$x, y \in \mathbb{R}, d \in \{0, 1\},$	M is a large constant parameter

The increase in modeling strength comes at a high cost. The introduction of

integer variables destroys the convexity of the solution space that makes LP problems relatively easy to solve.

A relaxed IP is the exact same model only the integer requirement is ignored. This implicitly transforms every relaxed IP or MIP to an LP as the only difference is the integer requirement.

The onerous difficulty in IP or MIP is that even if we have a solution to the relaxed IP or MIP problem, which can be found by LP solvers, we might be arbitrarily far from the optimal integer solution. A simple example that exhibits the problem:

$$\begin{aligned} \text{Minimize} \quad & z = x + y \\ \text{Subject to} \quad & x \geq 0.1 \\ & y - 200x \geq -20 \\ \text{Where} \quad & x, y \in \mathbb{N} \end{aligned}$$

The relaxed version where $x, y \in \mathbb{R}$ has the obvious solution $x = 0.1, y = 0, z = 0.1$, whereas the integer version stated above has the optimal solution $x = 1, y = 180, z = 181$. Considering problems with thousands of constraints this is a devastating distance.

The increased complexity has spawned a multitude of solution processes both exact and heuristic. Examples of exact generic approaches are Dantzig-Wolfe decomposition, Bender's decomposition and Branch and Cut all discussed in [Martin, 1999]. Within networks the algorithms by Dijkstra, Bellman and Ford and Floyd and Warshall are problem specific exact algorithms.

The plethora of heuristics developed and modified is huge, but they can be classified by the metaheuristic they are based upon with some few exceptions. Simulated annealing [Kirkpatrick et al., 1983], genetic algorithms, local branching and method of successive averages are a few examples of metaheuristics.

Considering ABIT, the increased modeling strength is much needed. On the other hand the solution methods may prove too ineffective.

3.1.3.3 Quadratic Programming

Quadratic Programming (QP) is an extension to LP where multiplicative terms between variables is possible.

This multiplicativity can then be used to express the complex interactions in the model. The decisive issue is that there is a very strict requirement on the

constraint matrix. Introducing multiplicativity in the formulation requires other mathematical methods to find and guarantee a solution. These methods have only been proven or found usable for very specific types of constraint matrices.

> Excerpt from Wikipedia.com on Quadratic Programming: For positive definite Q, the ellipsoid method solves the problem in polynomial time [Kozlov et al., 1980]. If, on the other hand, Q is negative definite, then the problem is NP-hard [Sahni, 1974]. In fact, even if Q has only one negative eigenvalue, the problem is NP-hard [Pardalos and Vavasis, 1991].

Some realistic problems have constraint matrices of these types, but it is generally accepted that even simplistic road networks do not. Furthermore the inseparable parts used for traffic modelling are more complicated than multiplicativity.

3.1.3.4 Mixed Complementarity Problems

Many real problems contain complementary constraints, i.e. pairs of constraints where at least one must be fulfilled. These are impossible to handle in LP and cumbersome in IP.

The solution is the introduction of the complementarity problem in [Isac, 1992]. The complimentarity operator $\perp$ is here used to formulate the complementing constraints. The meaning of $A \perp B$ is thus that constraint A or B must fulfilled. This means that if A is violated B must be fulfilled and vice versa. The requirement is not disjunctive as both A and B can be fulfilled simultaneously.

To show the modeling strength of Mixed Complementarity Problems (MCP) we will use Wardrop's first equilibrium as an example. It states that in this equilibrium no individual can reduce his perceived cost by changing routes.

From a different perspective this is the same as all used routes from one destination to another must be of equal perceived cost. Which again can be reformulated to: "only routes with the least perceived costs are traversed by vehicles"

Considering only one origin/destination pair the model simply becomes:

$$|p(x_a, x_b)_i| = 0 \perp f(p(x_a, x_b)_i) = f(\widehat{p}(x_a, x_b)) \qquad \forall i$$
$$|P(x_a, x_b)| \geq D_{x_a, x_b}$$

Where $f(\cdot)$ is the cost function and otherwise the notation from section 2 is

used. D_{x_a,x_b} is the demand of vehicles from the origin to the destination.

The definition is similar to the complementary slackness conditions of LP problems.

Some problems are easily represented as Complementarity Problems (CP). However the strength in formulation is encumbered by demanding solution methods. Mainly heuristics are applied today and quality of the solution is thus based on the allowed solution time.

If we intend to achieve Wardrop's first equilibrium the problem is easily formulated in MCP, but the social equilibrium is easier to represent in LP or IP.

Furthermore, according to [Friesz and Shah, 2001] the number of non linear complementarity constraints to represent the equilibrium are finite, but these are equalities. This is difficult as non linear equalities complicate any solution method that relies on moving in the search space as the moves are highly restricted.

The four approaches described so far are directly within the field of Mathematical Programming whereas the two last are inspired from different fields.

3.1.3.5 Variational Inequality Problems

Variational Inequality Problems (VIP) is the microeconomic approach to equilibrium problems. [Friesz et al., 1989] contains a more thorough discussion of this.

According to the article it is primarily based on the solution of a chain of states, which then reach a stable state either through tatonnement or disequilibria. Disequilibria are a set of states provably not in equilibrium which are iteratively modified by tatonnement until equilibrium is obtained. Tatonnement is alterations in the solution that attempt to establish equilibrium.

The general perspective is based on supply and demand from economics.

The supply with respect to traffic assignment is the capacity offered by the infrastructure. The demand is the requested trips by the populace. Urban network flows are then modeled into this through advanced delay functions and utility functions. These represent the complicated interactions on sections and in intersections. The approach offers extended versatility and modeling capacity, but is computationally demanding. The VIP approach is primarily aimed at

predictive dynamic transport assignment. This is used to, in greater detail, predict the reaction of the commuting populace to changes in the infrastructure.

As stated in [Friesz and Shah, 2001] the number of constraints representing user equilibrium is infinite and solving a VIP thus requires a cumbersome constraint accumulation scheme. Evidently it is not possible to individually represent an infinite number of constraints in a computer with finite memory. The solution is to start with a set of the constraints and then iteration by iteration find new violated constraints until the equílibrium is found.

Another drawback is the necessary preparations to use these models, which includes both crucial and time consuming calibration. The model requires a very representative utility function and not just the immediately measurable values of the network.

The iterational nature of stable state search may be far too time consuming to solve the entire problem. Recent research in [Patriksson, 2004] however shows that VIP can be used to find gradients in a running solution, which may enhance the tatonnement process.

The recent publication [Friesz and Shah, 2001] formulates a different and interesting disequilibrial network design which may prove efficient, but still requires significant theoretical development before it may be applicable.

3.1.3.6 Fluid Dynamics

The essence in Fluid Dynamics (FD) is to perceive the volume of vehicles as a fluid sieving through the network. The solution methods are highly evolved as FD has a long tradition starting from the much referenced article on kinematic wave theory [Lighthill and Whitham, 1955].

The transformation from road network to kinematic wave model is not without complications and much effort has been put into finding good aggregation methods. The main issue is that even though a lot of vehicles can be conceived as a fluid, the interaction at junctions cannot. The non-collision criteria combined with safety distances and headway makes the models extremely complicated or excessively inexact.

Beyond the modeling difficulties, the kinematic wave models do not yield a path flow distribution, which is what we need. Instead these yield a section flow distribution. This means that we have to find a path distribution given a flow distribution. Even though easier than finding a path distribution directly

it presents significant difficulties which effectively will eliminate the advantages in these models.

3.1.4 Simulation

As the power of computers has increased so has the applicability of simulational approaches to estimating flow distribution. Simulation offers an extreme modeling flexibility as the rules of the simulation does not have to fulfill a host of mathematical requirements.

Details that are close to impossible to represent in a mathematical model even at an aggregated level can easily be included in the rules of the simulation.

The microscopic capabilities of simulation is thus unprecedented in TS.

The drawback, however, is the often staggering strain on system resources when performing the simulation. In our case, the most evident problem is that the result of the simulation is not known until it is completed. It is not easy to extract usable early information and thus ensure quick response times. Furthermore it requires a separate simulation run for each alternative routing plan.

Another reason for not using the simulation approach is that scalability is poor. The extreme detail level possible is not needed as we can not apply the information properly. Reducing the level of detail brings us to a point where the strengths of an analytical model and solution is preferable to coarse grained simulation.

3.2 Solution methods

This section describes the evolution of the different solution approaches to the TAP and DTA.

3.2.1 Frank-Wolfe algorithm

The Frank-Wolfe algorithm was first published by [LeBlanc et al., 1975] and was the first known efficient approach to solving the TAP. The actual formulation of the generic Frank-Wolfe algorithm was almost two decades earlier in

[Frank and Wolfe, 1956]. [LeBlanc et al., 1975] was the first group of scientist to actually apply Frank and Wolfe's approach to the TAP.

The prerequisites of the Frank-Wolfe algorithm is a problem with a set of continous variables. It must obviously be possible to estimate the objective function at every feasible point. Dependent on properties of the cost function the Frank-Wolfe algorithm guarantees to find either a local or a global optimum[1].

The variables of the TAP is the number of vehicles on each path and the values should thus be considered as integer. However, this is an insignificant addition of uncertainty given the statistical approximations applied to calibrate the cost function. The integer requirement on the variables is thus dropped. The cost functions for the individual sections are defined for the non negative domain $\mathbb{R}_0$. Most cost functions, and especially those used in relation to TAP, resulting in a realistic objective function is defined for every non negative set of variables.

As the prerequisites are fulfilled [LeBlanc et al., 1975] uses a specific model for an adaptation of Frank-Wolfe's generic approach.

Even though this approach was a break through for TAP it suffered from getting caught in Steifel's cage. When caught in Steifel's cage the algorithm converged unexpectedly slow and it was quickly realized that the problem was caused by residual flows on inefficient paths. However as the original formulation did not include the entire set of generated paths this was not solved until much later.

The publication [Ran et al., 2002] presents an interesting approach to the DTA, where the commuters are divided into several classes dependent on the level of information that they have available. The authors show that different cost functions can be applied for the different levels of knowledge and propose an adopted Frank-Wolfe method using Method of Successive Averages (MSA) to solve their version of DTA. A more interesting mathematical contribution in the paper is the possibility of adding time and space to the standard TAP formulation without increasing the number of variables significantly.

3.2.2 Newton Method

The drawback of the original approach presented in [LeBlanc et al., 1975] was the inefficient descent direction. This was primarily caused by only considering the first order derivative of the cost function. This section briefly describes the underlying theory for the interested reader.

[1] Maximum or minimum dependent on the original problem

To introduce the more complicated Newton methods for unconstrained optimization we need to cover some mathematical concepts.

The full Taylor expansion of an expandable function is very precise close to the origin of the expansion and polynomially approximates the function at any other point. However the full Taylor expansion of complicated functions is intractable to apply. For the TAP a truncated taylor expansion is used to find the new minimum. Due to the computational limitations when [LeBlanc et al., 1975] published his paper the applied Taylor expansion was truncated to only considering the first derivative of the cost function – called the first order Taylor expansion.

The approximation error even close to origin caused the iterations to be caught in Steifel's cage. Considering a cost function like $f(x) = f_0 \cdot (1 + 0.15(x/100)^4)$ shows that there was another 3 non-zero derivatives that were ignored. The extention of this to a less truncated Taylor expansion – the second order Taylor expansion or the quadratic model – introduces the Hessian matrix to a function.

The quadratic model or second order taylor expansion of a function can be given as:

$$q(h) = f(x) + h^\top f'(x) + \tfrac{1}{2} h^\top f''(x) h \tag{3.2}$$

The Hessian is here stated as $f''(x)$.

In our case we seek an extremum of the function. Analytical mathematics indicates that for the function q we can only find an extremum at a stationary point. A point where the first derivative of q is 0. This consequently means that we only have to inspect the point or points where $f'(x) + f''(h) = 0$ is fulfilled to find the specific extreme point we want – the minimum. Any point that has the same functional value as the minimum for the function is called a minimizer for the function.

We thus solve the equation $f''(x)h_n = -f'(x)$. If the Hessian of f is positive semidefinite – all the eigenvalues of the Hessian are non-negative – the point h_n is a minimizer for q. As we have minimized the approximation of f Newtons method moves to the new point $x = x + h_n$ and calculates the new quadratic model and again finds the minimum. This is iteratively repeated until a convergence criteria is fulfilled. A convergence criteria could be that the found minimum only changes a specific amount per iteration. It could also be a required gap distance to a known lower bound for the minimum.

The following promising property of the Newton method is valid: "If an iterate x is sufficiently close to a local minimizer x^* and $f''(x^*)$ is positive definite, then Newton's method is well defined in all the following steps, and it converges

quadratically towards x^*"

However there are some disadvantages in the pure Newton method:

1. It is not globally convergent, which means that if we do not start with a proper initial point the minimizer will never be found
2. If the Hessian is not positive definite it might converge to a saddlepoint, which is neither minimizer or maximizer.
3. Requires the second order derivatives of the original function.
4. Specifically in our case the Hessian is huge for the larger difficult instanses using up to 400.000 variables. The full Hessian is thus 160.000.000.000 entries, which is unlikely to be efficiently possible on any computer today.

However the software we use in our computational experiments applies a Quasi-Newton Method (QNM). Essentially the exact Hessian is not used, but instead approximated through advanced updating schemes like the Davidon-Fletcher-Powell (DFP) or the more popular Broyden-Fletcher-Goldfarb-Shanno (BFGS) update method. Both are well known in relation to QNMs. Furthermore the Hessian is not stored in full. The Cholesky factorization of the Hessian is used instead. For the sparse matrices caused by our cost function this means that the problem size we can solve is increased dramatically.

We have here given a overview of the significant development within unconstrained optimization, which is the theoretical base for our solver. We consider it beyond the scope of this thesis to cover advanced QNM updating methods and Cholesky factorization. The article by [Bierlaire and Crittin, 2006] references some essential papers and books related to our problem.

3.3 Application of mathematical theory

For many years it was known to be intractable to represent all paths in a model and many TS researches inconsiderably abandoned any model with explicit path enumeration. I can only contribute this detour from best OR practice to the complications involved in Dantzig-Wolfe decomposition, which has been known since published in [Dantzig and Wolfe, 1961]

As we shall see there has been a significant delay between the establishment of new mathematics or solution methods and the application of these to TS. In

the case of explicit path enumeration this has significantly reduced the possible applications of TS for decades.

[Friesz and Shah, 2001] directly states that artificial intelligence – also called heuristics – is an efficient way of solving these problems. What they do not mention is the loss of the mathematical information which is available in the exact models. The idea in a heuristic is that some of the difficult mathematics or computations are substituted by easier approximating aggregations. This obviously reduces the calculation time per iteration, but usually requires significantly more iterations. The analytical properties gives by the complex mathematics is subsequently lost due to the simplifications. However some problems are analytically intractable, but approximate solutions can be found by applying heuristics.

Recently [Bierlaire and Crittin, 2006] has published interesting results on solving large-scale fixed-point problems and systems of non-linear equations. They present a generalization of secant methods, and uses several iterates to generate linear approximations. The method belongs to the Quasi-Newton family of methods, but their approach is matrix free, thus allowing them to solve large-scale systems of equations. Furthermore the method is not dependent on the existence of derivatives of the equations and there is no particular assumption on the problem structure or the problems Jacobian. Their approach is not directly applicable to our problem, but in time it will be very interesting to see the impact of their theoretically advanced approach. Hopefully it will take less than two decades for their method to be accepted in mainstream traffic research.

3.4 Applied practice

This section will cover the evolution of traffic information systems from a practical point of view. The idea is to cover the technology that has been developed to increase the information flow to the drivers. In short it will go from newspaper information to traffic radio and it's significance from early in the congestion era until today and the newer systems based on the increased information flow.

Traffic information has been distributed for ages. It is of such great importance that several species other humans are dependent on their ability to relay information related to paths or locations. Ants can lay down a scent track, which other ants follow and bees are capable of informing the rest of the worker bees in the hive by specific movements called dancing.

Since motorized vehicular traffic was commonly available it has become increas-

ingly important to be aware of the status of the infrastructure. In the early days of poor road quality it was important to know which roads had broken down or was being repaired. As newspaper was the main media used for information distribution the commuting people were informed in advance through these. The nature of a newspaper however requires the reader both to be able to read and to actually find the specific column of interest.

As radio was made generally available a new powerfull media for relaying realtime information was available. The requirement was now reduced to only listening to the radio. Most importantly it was possible to communicate directly to all listening drivers from a central position. Before the invention of cellular phones the information was gathered through aerial survailance in helicopters or by local reporters. It was important for drivers to know whether there was a better alternative. Even though the information broadcasted through radio was crude it was sufficient to ease the everyday commuting of traffic.

Since the development and acceptance of the Radio Data System (RDS) standard increasingly more car radios are capable of automatically tuning in on traffic information when broadcasted. This alleviates the necessity for drivers to be listinging to a specific radio channel.

The first message signs realized as Variable Message Signs (VMS) was fairly simple and could only be changed manually. This was used for lane closures and cautioning commuters in specific situations. The evolvement of display technology and communication technologies has enabled the signs to be changed dynamically, hence Dynamic Traffic Signs (DTS). The two terms are, however, used interchangably today regardless of the inuended difference in variable and dynamic. In practice the signs are essential as the information displayed is visible to every driver. Further more only drivers presumed interested in the information can be informed. This effectively reduces the amount of information that is given to the driver, thus minimizing the attention necessary to utilize it. The information distributed through these signs can be far more specific. Behind the DTS or VMS is Intelligent Vehicle-Highway Systems (IVHS), Advanced Traffic Information System (ATIS), Vehicle Information and Communication System (VICS) or another theoretical approach to decide what to communicate to the drivers.

Over the last years the in-vehicle combination of Geographic Information Systems (GIS) and Global Positioning System (GPS), known as route planners, has been increasingly used by commuters. However, as mentioned earlier, they only increase the static information level of the driver. To everyday commuters this is almost an insignificant increase as the driver travels almost the same route everyday. However to drivers that visit many different destinations and from different origins this is crucial. The British motorists association has been

pushing legislation requiring route planners in all rental vehicles to reduce the number of kilometers driven unintendedly. This would also provide better safety as given proper information presentation the locally unexperienced driver would be sure to reach his or hers distination efficiently.

Traffic Message Channel (TMC) is a manual dynamic update mechanism that allows the dynamic information in the route planner to be updated with real time information. The recent acceptance of the TMC standard has motivated hardware and software producers to include this partially dynamic enhancement of their product. However the information usually distributed is mainly general information on queuing or road works. It is unclear how detailed information is or can be.

For local areas information obtaining and distribution has been applied practice for quite a while. The list below describes some of the primary ITS initiatives:

Intelligent Vehicle-Highway Systems (IVHS) is the general name for initiatives taken after the US congress passed the Intermodal Surface Transportation Efficiency Act (ISTEA) in 1991. One of the main arguments for the ISTEA was a possible conservation of 20 billion USD by alleviating and bypassing congestion. In 1994 the US Office of Technology completed a study showing that 100 Billion USD in productivity was lost every year due to transportation difficulties. IVHS encompasses many initiaves such as reducing polution, improving car safety and increased real time information for drivers. The real time systems fit in the category Advanced Traffic Management Systems (ATMS), which are in effect numerous places in the states. Information is gathered manually or to some extent automatically and processed at a central information site. This processing is mainly based on manual inspection of a situation or state of the infrastructure. The outcome is dedicated information on DTS or traffic radio. Later and more advanced installments allow for dynamic assignment of red-green periods or lane closures. [Florida Department of Transportation, 2006] and [TCC, 1999] give further information on IVHS/ATMS. [MTO, 1999] is an american ATMS site and the danish site [DRD, 2004] would be considered an ATMS project.

PROMETHEUS an on board traffic safety enhancement project under the EUREKA program. The scope of the collection and application of information is only the individual vehicle. The PROLAB2 [Rombaut, 1995] vehicle presented in 1994 was an example of the increased possibilities in Pervasive Computing (PC) applied to vehicular safety.

General European Road Data Information Exchange Network also named (DRIVE) is an initiative program to standardize the Electronic Data

Interchange (EDI) of traffic information. It is unclear if TMC is a result from DRIVE projects.

Vehicle Information and Communication System (VICS) is the nationwide application of traffic information distribution in Japan. An elaborate interactive demonstration of the project can be found on this website [Information and Center, 2006]

3.5 Summary

The size and type of the problems that can be considered today have been equally dependent on computational power as well as significant methodological advancements. The discrete choice modelling approach considering the rational human – homo economicus – has been crucial in the development of representative theoretical results. The cost functions constructed from this research has usually been aimed at Wardrops first equilibrium – the user equilibrium. Wardrops second equilibrium – the social equilibrium – has however not been considered much in litterature. It is of theoretical interest to find the social equilibrium for a given network, but in practice only the user equilibrium is encountered. The application of ABIT is real-time operational traffic planning. Introduction of ABIT should ensure faster reestablishment of either of the equilibria in the case of disruptions. This requires sufficiently good solutions fast. Combining this with different modelling and solution methods indicates the the generic Frank-Wolfe method (FW) approach can be augmented to function efficiently even with inseparable cost functions. This is exactly what we will be focusing on in the following chapters. The inseparable cost functions are needed to increase the representativeness of a a model. The interesting work [Bierlaire and Crittin, 2006] combines many newer advanced mathematical findings to a new augmented solver.

In practice the evolution of traffic information distribution has evolved to increase the onsite information as well as filtering irrelevant information for the driver. Moving from newspaper to Radio Data System (RDS) was a significant evolutionary step. The next step was the Variable Message Signs (VMS) or Dynamic Traffic Signs (DTS), which gives local information to everyone passing a specific sign. The introduction of TMC allows route planning devices – like GPS navigators – to receive dynamic information on different types of disruptions. Thus enabling real-time information filtering and processing only reflected to the driver by a change in the proposed route.

The next chapter on theory covers details considered for ABIT.

Chapter 4

Theory

In this chapter we will focus solely on the theoretical basis and further development of the software produced with the Quasi-Newton Method for TAP with inseparable cost function found in appendix D. The primary difference to most research today is the application of a more versatile solver, that guarantees convergence even with the more realistic inseparable cost functions.

4.1 Our model

As described in the previous chapter numerous models for Traffic Assignment Problem (TAP) has been developed. For the Quasi-Newton Method for TAP with inseparable cost function we adopt a very common formulation for the overall problem:

Minimize	$\sum_{n,i,j} \int_0^f T'_n(x)dx$	User equilibrium
Subject to:	$\sum_k p_{ijk} \geq D_{ij}, \forall i, j$	Demand satisfaction constraints.
	$p_{ijk} \geq 0$	Paths flows non negative

This model is the Bechmann transform for the TAP. The Bechmann transformation introduces the integral in the objective function and ensures that the equilibrium found is the user equilibrium and not the social equilibrium. f is

the combined flow on all paths using a specific section. This aggregation is important if the cost function is non linear. In our case and in many other publications the cost function is non linear. However, we also introduce inseparability in the cost function, which is different to most published articles on related subjects.

4.1.1 Cost function

Our cost function is based on the mostly used formulation:

$$T_{\overrightarrow{x_a x_b}}(|\overrightarrow{x_a x_b}|) = fft_{\overrightarrow{x_a x_b}} \cdot \left(1 + \alpha \frac{|\overrightarrow{x_a x_b}|}{\lceil \overrightarrow{x_a x_b} \rceil}\right)^{\beta}$$

where $fft_{\overrightarrow{x_a x_b}}$ is the free flow time – the traversal time for a vehicle if no other interfering traffic is present – for the section. α and β are calibration parameters along with the theoretical capacity $\lceil \overrightarrow{x_a x_b} \rceil$. The above formulation is widely adopted as it is separable and thus guarantees convergence for even simplistic solvers.

As mentioned in the article in appendix D on page 127 enforcing this separability reduces the representativeness of the cost function. As the solver we have constructed is proposed to be independent of separability we can add a significant real world complexity to the cost function, albeit at the cost of longer solution times.

Many others have already addressed congestion spill back. Here we only consider yield rules in regular traffic. We have not found any litterature attempting to include this into a TAP solver previously. This is interesting as constructing a proper function representing the yield impact has been researched for several years. In traffic litterature a function that represents the waiting time for the yielding traffic is called a gap acceptancy function (GAF). Our guess is that it has been excluded from the requirements of the solvers exactly because of the introduced inseparability. Given our solver and the theoretically stronger solution method proposed in [Bierlaire and Crittin, 2006] this may soon be a less significant issue.

Our cost function is thus extended with a generic GAF:

$$\left(e^{f_n f_m} - f_n f_m - 1\right) f_n^{-1}$$

where f_n represents the flow that has to yield and f_m represents the opposing flow.

The combined cost function can then be written as:

$$T_{\overrightarrow{od}}(|\overrightarrow{od}|) = fft_{\overrightarrow{od}} \left(1 + \alpha \frac{|\overrightarrow{od}|}{\left\lceil \overrightarrow{od} \right\rceil \cdot \gamma_{\overrightarrow{od}}} \right)^{\beta} + \sum_{0}^{m} y_{nm} \left(e^{f_n f_m} - f_n f_m - 1 \right) f_n^{-1}$$

where the binary variable y_{nm} determines whether the flow on section n has to yeild for the flow on section m.

The GAF that we use is simplified, but does contain the exponential function, which is widely used in different GAF functions.

Congestion spillback can obviously also be approximated by representative functions and dependencies. As described later our solver has only one simple mathematical requirement. An effect of this is that many different approximations previously considered unusable are possible.

4.2 Basic Solution Method

Our solver shown in figure 4.1 is constructed much like the traditional approaches to the TAP. The significant difference to other approaches is that in step 5 we apply a true Quasi Newton Method. The effect of this is that we can guarantee a good convergence rate even for inseparable cost functions. Evidently our Quasi Newton Method is more cumbersome than the traditional Pseudo Newton approaches and thus requires more time. However, the solutions found are more representative as significant dependencies in the network can be modelled.

4.2.1 Shortest paths

Finding the shortest path in a network is a well researched topic and entire books on the subject is readily available to the interested reader. The algorithm we apply through the GOBLIN software is an optimized version of a shortests path algorithm. It has been implemented as a FIFO label correcting algorithm, which is a practical improvement on the Bellman-Ford algorithm. The experimental results from the article show that only 0.5% of the running time is spent on generating paths. It might be interesting to specialize the path generation to dynamic path costs. This could possible reduce the number of flow redistributions, which have been proven to be the far most timeconsuming part of the solution process.

1. Set all flows on all sections to zero and the size of the paths pool to zero.
2. Find all pairs shortests paths.
3. Add all paths found not already in the pool to the pool.
4. If no new paths were added to the pool go to step 8.
5. Distribute the demand flow with the Quasi-Newton Method (QNM) with the path pool.
6. Set the flows on all sections according to the found distribution.
7. Go to step 2.
8. The path pool and the distribution found by the QNM solver is the best solution for the predefined convergence criterion.

Above step 5 is handled by MINOS and step 2 is done through GOBLIN. Both are available software packages. The MINOS software is however a commercial package and requires a license.

Figure 4.1: Construction of our solver

4.2.2 Flow distribution

The flow distribution performed in our solver by MINOS is an efficient Fortran imeplementation of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) QNM. However to achieve good scalability Cholesky Factorization of the Hessian is used to reduce the memory consumption. MINOS has with our solver redistributed the flow in a system consistings of up to 500.000 variables and almost 900.000 constraints.

4.2.3 Dynamics

So far our work has only extended the traditional static TAP to include include inseparable cost functions. The previously mentioned dynamic extension to the TAP, the Dynamic Traffic Assignment (DTA) has not been covered.

The DTA included the possibility of demand changing over time. A possible approach is thus to add an integration over time to the objective function. Another previously used approach is by expanding the network represenation

with replicas of nodes and sections for different time periods.

The addition of an integral will only make a difference if the existing objective function does not have a computable integral. If the integral is efficiently computable and locally polynomial it should make no difference to our MINOS based flow distribution. This is mainly caused by the fact that we are not bound to separable cost or objective functions. Only the cost function calculations might be more expensive in this part of the algorithm. However the path generation algorithm has to be far more detailed than the present static one. This might lead to worse scalability as the size of the network searched for improving paths increases exponentially with the required dynamic detail. Using network expansion by replication, however, will use the same cost function, but still require the increased precision in the path generation.

Using replication seems to be the simplest approach, but adding the integration to the cost function increases the representativeness. The choice however is far to complicated to simply decide from a deduction. Both alternatives has to be implemented and tested to make a qualified decision. The following two sections on time and space and forecasting are included as a discussion on further theoretical aspects relevant for Agent Based Individual Traffic guidance (ABIT).

4.3 Time and space

The versatility gained by considering inseparable cost functions allows our type of solver to be used for more than just small or large areas. By using hierarichal aggregation it is possible to represent even very large networks and get long distance routing information that is locally specific.

The instances we solve in Quasi-Newton Method for TAP with inseparable cost functiondo not give a clear picture of the small scale behavior of the solver. It is, however, clear that considering 3000 sections and 1000 nodes is not presently possible. Considering 200 sections and 50 nodes allows us to use non aggregated information in the immediate travelling direction and aggregate subregions into individual nodes surrounding these. At greater distance, regions or groups of regions can be aggregated into nodes as well. Combining several super regions into one node is viable for very long range route planning. Figure 4.2 on the next page depicts an example of the aggregation.

The dynamic aggregation of the considered network reduces the precision of the route. The solution found in the aggregated network can however be shown on a regular map. Selecting an exact route based on the overall directions found in

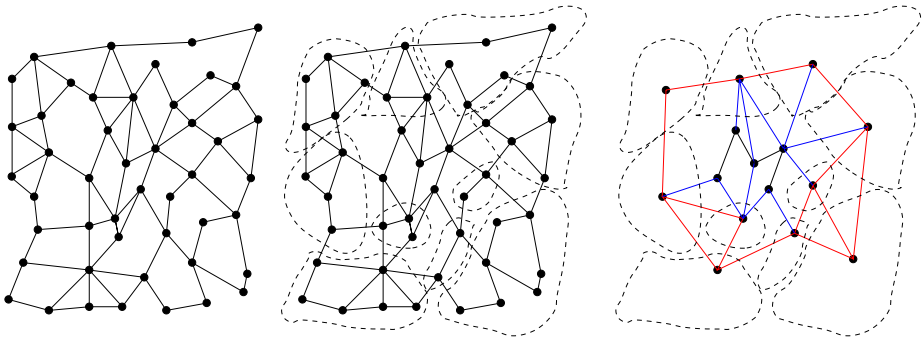


Figure 4.2: Aggregation example

the aggregated network makes it possible to present an entire expected route. The increased information and dynamic state updates in ABIT make in-transit route changes more likely. The imprecision in the aggregated solution is thus to be expected.

The actual process of spatial aggregation proposed above is a complicated task that requires efficient methods and possibly timeconsuming calibration. This process however can be done in advance. When a disruption is detected only the vehicle distribution part considered in Quasi-Newton Method for TAP with inseparable cost function is thus necessary. As the solver we apply is capable of handling inseparable cost functions the aggregated representation can contain more complicated flow relationships than with the traditional approaches.

The long range route guidance, the individual alleviation described in section A.3.2 is thus handled by properly aggregating the entire network.

The immediate dispersion however does not need to consider the entire network as the immediate interest is to avoid escalation of the Level of Influence (LoI). Finding the maximum throughput of an area is consequently of great interest. Once the disruption is detected a small area is considered. If the maximum throughput of the area is sufficient for redistribution of the impeded flows only the small area is considered in the flow redistribution. If the maximum throughput is insufficient the area is extended until the maximum throughput is sufficient. We have not found any litterature on the maximum flow problem with inseparable non-linear cost functions. Complexity and runtime are thus unknown for solving such a problem.

4.4 Forecasting

Forecasting is well known in relation to wind and weather. Forecasts on traffic are less common. One of the reasons for this is the unpredictability of autonomous vehicular traffic. Concluding that a major road work will result in traffic problems in some region is obvious and a simple forecast. Actually increasing the precision of the forecast to which sections and to what extent the problems will be is far more demanding. The major obstacle in this direction however is not the theory on which to found a forecast or prediction, but the detail and amount of information necessary.

To make a reliable forecast requires a properly calibrated correct statistical model. This in turn requires detailed information on the real world instances and historical data for previous encounters. This is practically impossible today as the only way to accomplish this is by manual or mostly manual means. Within the near future Global Positioning System (GPS) and other systems aimed at road pricing will be capable of automating the dataacquisition. Once ABIT or another Pervasive Traffic Intelligence (PTI) is deployed the dataacquisition can be made fully automatic.

The application of advanced statistics and forecasting will thus be possible. As mentioned in appendix A this is an important part of making a reliable system.

Chapter 5

Papers included in the thesis

The work on Agent Based Individual Traffic guidance (ABIT) has led to several different papers. Each of these are based on different perspectives on the topic. The first paper is introductory, the second cover future perspectives, the third considers measuring and combining measures. The fourth covers the impact and effect of introducing new theory into practice, The fifth contains results that support or illustrate findings or further research topics. The final paper covers the highly essential multi dimensional random number sequences that must be available for calibration and forecasting mentioned repeatedly through this thesis.

In this chapter we will shortly describe the relevancy and findings of the six papers in the appendix.

5.1 Agent Based Individual Traffic guidance

The first paper in appendix A was written and published in 2004 on ABIT and it contains the initial consideration on Pervasive Traffic Intelligence (PTI). It describes the currently available information channels for traffic. The spatial and temporal localities are introduced here along with the Level of Influence (LoI).

Furthermore the operation of the ABIT system is described as two levels.

Immediate dispersion considering only the disruption and finding the best diversion or bending pattern, and

Individual alleviation considering the longer distances aimed at providing best possible routes for individual commuters.

This distinction is based on the different practical and theoretical requirements for the two. The immediate dispersion can use immediate and local information. The efficiency can be increased by using short term forecasting. However it requires fast and efficient solutions as the reaction time of the system is crucial.

Individual alleviation requires information from distant areas and thus requires advanced forecasting and aggregation to give stable paths for the commuters. The issue here is not necessarily immediate response to a disruption, but instead qualified response to the commuter. The time alloted for a solution is thus more, but the robustness of the solution must be better.

5.2 ABIT measuring

The paper in appendix C was written subsequently to cover the considerations on the characteristics usable in differentiating alternate routes and optimization criteria.

The different characteristica are discussed in relation to applicability and usage in relation to ABIT. It was decided to consider time and distance as the main constituents for the cost function. Articles on route selection using discrete choice modelling point to these as most significant for the drivers. For us time is the difficult dimension as it introduces the difficult inseparability. Time is also cumbersome to model and forecast due to the dependency on the specific flow and types of vehicles on a given section.

The paper also addresses different objective constructs and three different objectives were selected. However, neither of these were applied in the later research as the intent of the Quasi-Newton Method for TAP with inseparable cost function paper was to compare the solution time to another efficient solver.

5.3 Agent Based Individual Traffic guidance 2

The second article accepted at the conference "Trafikdage" in 2006 is included as appendix B. This article covers our considerations on the perspectives of ABIT.

Recent years of increased impact of rush hour has increased both research and public interest in congestion. Over the last four years the publications relevant for this topic have increased significantly. The processing capabilities of the computers have along with advances within solution methods contributed to the increase in application of Intelligent Traffic Systems (ITS). Although no actual PTI system is running today many interesting systems are performing on the operational level of traffic planning.

The research on Pervasive Intelligence (PI) and Ambient Computing (AC) indicates that massively deployed intelligent systems are emerging. We thus believe that both the theoretical and practical requirements for ABIT are fulfilled within a few years.

The paper can be considered a long conclusion on the project in general and the expectations regarding the deployment of PTI variants.

5.4 Quasi-Newton Method for TAP with inseparable cost function

This theoretical paper on a specific experimental solver implementation is included as appendix D. It contains the considerations on an advanced cost function and the requirements to a solver capable of efficiently solving the problems. The paper also contains the findings from an implementation of the solver based on the software packages GOBLIN and MINOS. The findings give good indications on the abilities and efficiency of a solver capable of using an inseparable cost function. Inseparability in the cost function can also be used to increase the precision in the automated aggregation of the network both for immediate dispersion and individual alleviation.

The yield cost impact were illustrated for small instances and the effect of disruptions in the network on solution time was shown for a much larger instance. The solution times were also compared with an efficient solver expecting separable cost functions. Our solution implementation scales significantly worse, but the increased solution time is partially countered by increased precision in

the modelling capabilities for the experimental solver.

5.5 Further results

The paper in appendix E on results generated with the above solver contains the findings that support or illustrate expected situations or other relevant information. The effects of disruptions on a network is illustrated through examples. The diversion of traffic is expectedly similar to the bending curves depicted in figure A.9 on page 90. The idealized bending in the figure is approximated in relation to the heterogenous capacities on the sections.

5.6 Drawing a random number

Several topics related to optimization and number theory are relevant for traffic modeling. The paper is the result of a course assignment in a course on this topic related to generation of multidimensional random numbers based on the Halton sequences.

The motivation for the paper is the long and still ongoing discussion between researchers on different methods of generating multidimensional numbers. Although only indirectly applicable to this thesis and the work done so far, the findings are important when working with statistics.

Our findings are significant in determining how to generate multidimensional random sequences, which are required to estimate and calibrate even simplistic statisticals models. Of specific importance is the result showing that it is practically impossible to generate truly uncorrelated multidimensional Halton sequences of more than 40 dimensions. This result well become relevant once forecasting and other statistical approaches are applied in relation to ABIT.

Chapter 6

Conclusion

Traffic is increasing and the impact of the increased private commuting is resulting in congestion in almost every larger city. The topic is discussed regularily in the press and political forums. The costs associated with significant changes in the infrastructured are immense. This has for decades given funds and incentives for research and practitioners to address traffic related problems.

It is commonly accepted that to achieve as much as possible from the existing infrastructure the drivers have to be well informed of the real time situation elsewhere in the network. Traffic radio was the first real-time relaying of information. Even though we have found no direct documentation on the effect of traffic radio, the increased information available on disruptions in the network has reduced the travelling time for many commuters. The following development of Dynamic Traffic Signs (DTS) and Variable Message Signs (VMS) increased the local precision of the information. It is an important supplement to traffic radio as more specific information can be relayed to local commuters only. The recent development, deployment and commisioning of Traffic Message Channel (TMC) has made the information automatically available to the otherwise static Global Positioning System (GPS) route planning devices existing today.

However, all these are based on manual or automatic static information gathering as the individual route planning devices are incapable of transmitting information. In this thesis we assume that the onboard devices are capable of

such and can function as Pervasive Traffic Intelligence (PTI). PTI is Pervasive Intelligence (PI) used for operational traffic planning. The idea is that every Agent Based Individual Traffic guidance (ABIT) enabled vehicle can transmit real-time information to a Central Information System (CIS). The CIS then processes the information and redistributes this to the vehicles. The onboard device can then assist the driver by proposing expected best routes when given the information. This is much different from today's systems which at best, with TMC, rely on static or simplistically estimated information.

The intent of ABIT is to increase the information precision and guidance to the individual driver. The goal is to distribute the traffic to Wardrop's first equilibrium as fast as possible. Once the infrastructure is disrupted, the individual driver or TMC enabled route device have to guess the best new path. With the partially centralized information processing and route generation it is possible to guide the traffic to the new equilibrium efficiently. The time lost in suboptimal flowdistribution in a non-equilibrial situations can thus be minimized.

Although Wardrop's second equilibrium, the social equilbrium, is more efficient we find it unlikely that the autonomous driver will accept a slower route to speed up the transit of others.

To address the topic of a disruption we defined the spatial and temporal locality of a disruption. Furthermore the Level of Influence (LoI) of a disruption is also introduced to indicate the impact of disruption. Through scenarios on user imposed deviation, accidents, emergency vehicles and other causes, it is described how these characterstics can be used as the "what, when and how" of a disruption.

The ABIT system was also described as two different levels: the immediate dispersion and the individual alleviation. The immediate dispersion is used for the immediate short term redistribution of flows close to a disruption, whereas individual alleviation is aimed at longer distances. Individual alleviation is expected to require far more complicated forecasting and aggregation than the short ranged immediate dispersion.

The basic problem we consider is the Traffic Assignment Problem (TAP), which later evolved into Dynamic Traffic Assignment (DTA) with the inclusion of changing demand over time. Traffic assignment considers a cost function which is used to model the cost of traversing a given section based on a set of parameters. We chose time and distance as the primary parameters here.

Representing these requires models for both. Through discrete modeling and the perception of the commuting populace as homo economicus – the rational man – it is possible to devise a representative cost function. The difficulty is

time, which is dependent on the load of a section. We chose this cost function:

$$t(|\overrightarrow{od}|) = fft \left[\left(1 + \alpha \frac{|\overrightarrow{od}|}{\left|\overrightarrow{od}\right|} \right)^{\beta} \right] \tag{6.1}$$

which was proposed in [Bureau of Public Roads, 1964]. This cost function has been generally accepted as the best approach. The function is separable, which is an attractive mathematical property when solving the instances. The drawback however, was that the dependency between flow on different sections cannot be modelled.

The precision we need for ABIT is different than usual and we introduced the gap acceptancy function (GAF). This function represents the impact of yielding traffic passing through right-of-way traffic. This is exemplified in any light controlled intersection where cross turning traffic needs to yield for traffic going straight through the intersection. This introduces inseparability, which is a onerous mathematical property when solving the instances.

We have constructed a Frank-Wolfe method (FW) like approach using a true Quasi-Newton Method (QNM) through state-of-the-art software packages and compare the performance of our solver with the OBA-solver [Bar-Gera, 2002].

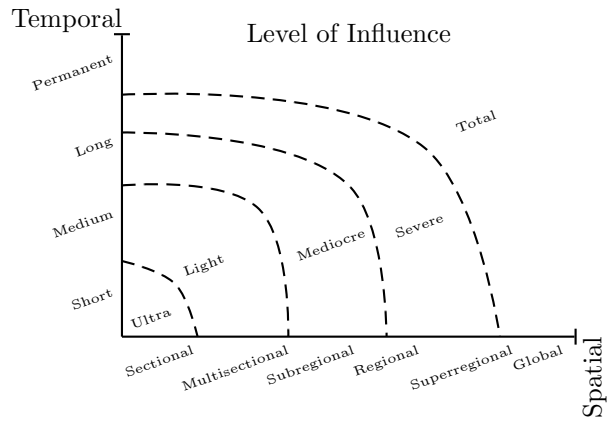
Our performance is expectedly slower than the separable solver. Recent research found in [Bierlaire and Crittin, 2006], however, shows that others are using even more advanced methods to address inseparable cost functions. Their reference to the recently defined Consistent Anticipatory Route Guidance (CARG) problem shows that present day research is evolving towards operational traffic planning usable for ABIT.

We expect the performance of inseparable solvers will increase significantly. They might even be able to compare with separable solvers on separable instances.

The experimental solver we constructed was significantly slower than state of the art separable solvers. With our solver we were able to show the impact of using a simplified gap acceptancy function. This impact becomes increasingly important for high precision operational traffic planning.

As processing power and communication methods are evolving constantly, we believe that ABIT is technically possible today.

We do not expect that neither practical nor theoretical difficulties will be hindering a possible deployment within a few years. The main difficulty is the



The dashed lines can be considered as iso-LoI lines

Figure 6.1: Classification guide

political and commercial interests in realization of any PTI, such as ABIT.

Whether the increased and correctly distributed information will have a significant impact is yet unknown. ABIT is inherently dependent on the decisions made by the autonomous driver and can therefore be a tool to make a difference.

Even though there are some cumbersome steps remaining I believe that ABIT or an equivalent PTI will be a viable aid to any disrupted situation in traffic in the near future. As traffic increases, and thus congestion, effective PTI systems, such as ABIT, will likely be of paramount importance for the commuting populace.

Main results

The thesis contributes to the field of PTI through the development and description of the central concepts in ABIT and their complex relationships. This included defining the disruption characteristics shown in figure 6.1 and applying these in selected scenarios. The interesting bending patterns depicted in figure 6.2 expected in the disrupted solutions were shown by applying our generic solver designed for inseparable cost functions. They show the importance of the introduced LoI of a disruption and the following alleviation. The working paper on measuring discusses many characterstics used in Discrete Choice modelling on route choice and their applicability in Operations Research (OR) models. The paper was concluded by a section on objective function constructs. Three distinctively different constructs were proposed from an OR perspective:

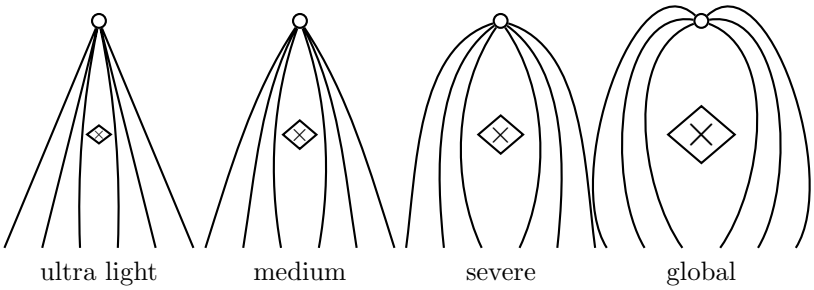


Figure 6.2: Level of Influence bending patterns

- The static time objective

$$\min\left\{\sum_{\forall i} t(p_i^{free})\right\}$$

- The dynamic time objective

$$\min\left\{\sum_{\forall i} t(p_i)\right\}$$

- The relative social time objective

$$\min\left\{\max_{\forall i} t(p_i)/t(p_i^{free})\right\}$$

The work on multidimensional random numbers based on Halton sequences exemplified the importance of understanding the generation of random sequences. Different generation methods were inspected and a set of guidelines for using Halton sequences were devised. It was also found that generating uncorrelated multidimensional random sequences of more than 100 dimensions is extremely cumbersome when using Halton sequences.

Topics for further results

During the work on this thesis the following topics were found interesting for further research:

- The hierarichal network aggregation described in chapter 4 on theory contains many different intricacies that must be addressed with great care. The issue of automated hierarichal network aggregation will be of great importance to the distributed nature of ABIT.
- The definition of LoI depends on very specific knowledge. Firstly, the reduced capacity of the disrupted section and secondly, the maximum possible through-put for a given area. While the new reduced capacity can be estimated or measured directly, the maximum through-put of an area is more complicated. Work on the maximum through-put of an area is essential for efficient estimation of the LoI of a disruption.
- We have already adressed the immediate dispersion to some extent, but the individual alleviation has not been considered in detail. Both aggregation and forecasting are crucial to the construction of an efficient and robust approach to individual alleviation.
- Solution methods and application of advanced specialized methods as in [Bierlaire and Crittin, 2006] related to CARG is interesting. The CARG problem contains many intricacies similar to the difficulties in PTI
- The QNM requires the gradient of as many variables as possible in each iteration. The acceptancy on inseparability through our approach enables researchers and practicioners to apply significantly different cost functions. Many of these might have easily computable values and derivatives. However, some cost functions will need efficient numerically approximations to their value or gradient for the solver to remain efficient.
- The fraction of ABIT enabled vehicles or necessary routing changes for alleviating a given LoI. The issue here is to examine the necessary redistribution and estimate the fraction of ABIT enabled vehicles in the network necessary to efficiently alleviate the disruption.
- The objective constructs proposed in ABIT measuring in appendix C require alterations to the solver not immediate possible. The new objective constructs may also exhibit characteristics that improves or impairs the convergence rate.

Bibliography

[Bar-Gera, 2002] Bar-Gera, H. (2002). Origin-based algorithm for the traffic assignment problem. *Transportation Science*, 36(4):398–417.

[Beckmann et al., 1956] Beckmann, M., McGuire, C., and Winsten, C. (1956). *Studies in the Economics of Transportation*. Yale University Press, New Haven, Conneticut.

[Bierlaire and Crittin, 2006] Bierlaire, M. and Crittin, F. (2006). Solving noisy, large-scale fixed-point problems and systems of nonlinear equations. *Transportation Science*, 40(1):44–63.

[Bonsall and Parry, 1990] Bonsall, P. and Parry, T. (1990). Drivers' requirements for route guidance. *Road Traffic Control, 1990., Third International Conference on*, pages 1 –5.

[Bronsted et al., 2005] Bronsted, J., Hansen, K. M., and Kristensen, L. M. (2005). An infrastructure for a traffic warning system. *Proceedings - International Conference on Pervasive Services, ICPS '05 and Proceedings - International Conference on Pervasive Services, ICPS '05*, 2005:136–145.

[Bureau of Public Roads, 1964] Bureau of Public Roads (1964). *Traffic Assignment Manual*. Urban Planning Division, US Deparment of Commerce, Washington, DC.

[Dantzig and Wolfe, 1961] Dantzig, G. and Wolfe, P. (1961). The decomposition algorithm for linear programs. *Econometrica*, 29:767–778.

[Department of Transport, 1985] Department of Transport (1985). *Traffic Appraisal Manual (TAM)*. HMSO, London, UK.

[Dong et al., 2006] Dong, J., Mahmassani, H. S., and Lu, C.-C. (2006). How reliable is this route? predictive travel time and reliability for anticipatory traveler information systems. *Transportation Research Record*, 1980:117–125.

[DRD, 2004] DRD (2004). Trim – traffic map [in danish]. Online by Danish Road Directorate. direct link http://www.trafikken.dk/wimpdoc.asp?page=document&objno=77436.

[Flagstad, 2006] Flagstad, T. (2006). Biltrafikken udleder stadig med kuldioxid. Nyt fra Danmarks Statistik 462, Statistics Denmark. Miljœ og energi, Trafik og miljœ 2006.

[Florida Department of Transportation, 2006] Florida Department of Transportation (2006). Florida's statewide its general consultant. Internet. direct link http://www.floridaits.com/.

[Forum, 2004] Forum, T. (2004). Tmcforum.com. Online by TMC Forum. direct link http://www.tmcforum.com/.

[Frank and Wolfe, 1956] Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110.

[Friesz et al., 1989] Friesz, T. L., Luque, J., Tobin, R. L., and Wie, B.-W. (1989). Dynamic network traffic assignment considered as a continuous time optimal control problem. *Operations Research*, 37(6):893–901.

[Friesz and Shah, 2001] Friesz, T. L. and Shah, S. (2001). An overview of nontraditional formulations of static and dynamic equilibrium network design. *Transportation Research Part B: Methodological*, 35(1):5–21.

[Hill and De Santis, 2002] Hill, A. and De Santis, M. (2002). Traffic monitoring application of cellular positioning technology: Proof of concept. Online pdf publication on Transport Canada's homepage. direct link http://www.tc.gc.ca/tdc/summary/13900/13936e.pdf.

[Information and Center, 2006] Information, V. and Center, C. S. (2006). Vics home page. Internet. direct link http://www.vics.or.jp/english/index.html.

[INFORMS, 2004] INFORMS (2004). Science of better website. Technical report, Institute for Operations Research and the Management Sciences, http://www.scienceofbetter.org/.

[Isac, 1992] Isac, G. (1992). *Lecture notes in Mathematics: Complementarity Problems.* Springer-Verlag.

[Jara-Diaz and Friesz, 1982] Jara-Diaz, S. R. and Friesz, T. L. (1982). Measuring the benefits derived from a transportation investment. *Transportation Research, Part B: Methodological*, 16B(1):57–77.

[Kirkpatrick et al., 1983] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 3.220(4598):671–680.

[Kozlov et al., 1980] Kozlov, M., Tarasov, S., and Khachiyan, L. (1980). The polynomial solvability of convex quadratic programming. *USSR Computational Mathematics and Mathematical Physics*, 20(5):223–228.

[LeBlanc et al., 1975] LeBlanc, L., Morlok, E., and Pierskalla, W. (1975). An efficient approach to solving the road network equilibrium traffic assignment problem. *Transportation Research*, 9(5):309–318.

[Lighthill and Whitham, 1955] Lighthill, M. J. and Whitham, G. B. (1955). On kinematic waves. ii. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 229(1178):317–345.

[Martin, 1999] Martin, R. K. (1999). *Large Scale Linear and Integer Optimization: A Unified Approach*. Kluwer Academic Publishers, 3rd printing edition.

[Meschini et al.,] Meschini, L., Bellei, G., Gentile, G., and Papola, N. An implicit path enumeration model and algorithm for dynamic traffic assignement with congestion spillback. Technical report, Dipartimento di Idraulica Trasporti e Strade, University of Rome 'La Sapienza' Italy. Publish year unknown.

[MTO, 1999] MTO (1999). Traffic and road information system (tris). Online by Ontario Ministry of Transportation. direct link http://www.mto.gov.on.ca/english/traveller/compass/systems/tris.htm.

[Ortúzar and Willumsen, 2001] Ortúzar, J. D. D. and Willumsen, L. G. (2001). *Modelling Transport*. John Wiley & Sons Inc., 3rd edition.

[Pang et al., 1999] Pang, G., Takabashi, K., Yokota, T., and Takenaga, H. (1999). Adaptive route selection for dynamic route guidance system based on fuzzy-neural approaches. *Vehicular Technology, IEEE Transactions on*, 48(6):2028 –2041.

[Pardalos and Vavasis, 1991] Pardalos, P. M. and Vavasis, S. A. (1991). Quadratic programming with one negative eigenvalue is np-hard. *Journal of Global Optimization*, 1(1):15–22.

[Patriksson, 2004] Patriksson, M. (2004). Sensitivity analysis of traffic equilibria. *Transportation Science*, 38(3):258–281.

[Pedersen, 2004] Pedersen, S., editor (2004). *Regionalt Trafikoverblik*. Hovedstadens Udviklingsråd, Gammel Kœge Landevej 3, DK2500 Valby, Denmark.

[Postrel, 2004] Postrel, V. (2004). Operation everything. The Boston Globe, June 27, 2004. direct link http://www.dynamist.com/articles-speeches/opeds/opresearch.html.

[Prager, 1954] Prager, W. (1954). Problems in traffic and transportation. *Proceedings of a Symposium in Operations Research in Business and Industry*.

[Psaraftis, 1995] Psaraftis, H. (1995). Dynamic vehicle routing: status and prospects. *Annals of Operations Research*, 61:143–164.

[Ran et al., 2002] Ran, B., Lee, D.-H., and Shin, M. S.-I. (2002). New algorithm for a multiclass dynamic traffic assignment model. *Journal of Transportation Engineering*, 128(4):323–335.

[Rich et al., 2003] Rich, J. H., Holm, J., and Nielsen, O. A. (2003). The helsingœr-helsinborg tunnel project: Atkins report. Technical report, Technical University of Denmark.

[Rombaut, 1995] Rombaut, M. (1995). Prolab2: A driving assistance system. *Mathematical and Computer Modelling*, 22(4-7):103–118.

[Sahni, 1974] Sahni, S. (1974). Computationally related problems. *SIAM Journal on Computing*, 3(4):262–279.

[Smock, 1962] Smock, R. (1962). An iterative assignment approach to capacity restraint on arterial networks. *Highway Research Board Bulletin*, 156(2):1–13.

[TCC, 1999] TCC (1999). Gta city status. Technical report, Toronto City Counsil, http://www.city.toronto.on.ca/. direct link http://www.city.toronto.on.ca/torontoplan/citystatus_5.htm.

[Wardrop, 1952] Wardrop, J. (1952). Some theoretical aspects of road traffic research. *Institution of Civil Engineers – Proceedings*, 1:325–362.

[Yagar, 1971] Yagar, S. (1971). Dynamic traffic assignment by individual path minimization and queuing. *Transportation Research*, 5(3):179–96.

Appendix A

Agent Based Individual Traffic guidance

Abstract

When working with traffic planning or guidance it is common practice to view the vehicles as a combined mass. From this models are employed to specify the vehicle supply and demand for each region. As the models are complex and the calculations are equally demanding the regions and the detail of the road network is aggregated. As a result the calculations reveal only what the mass of vehicles are doing and not what a single vehicle is doing.
This is the crucial difference to ABIT (Agent Based Individual Traffic guidance). ABIT is based on the fact that information on the destination of each vehicle can be obtained through cellular phone tracking or GPS systems. This information can then be used to provide individual traffic guidance as opposed to the mass information systems of today – dynamic road signs and traffic radio. The goal is to achieve better usage of road and time.
The main topic of the paper is the possibilities of using ABIT when disruptions occur (accidents, congestion, and roadwork). The discussion will be based on realistic case studies.

A.1 Traffic Today

The objective of this document is to describe different situations in the ABIT (Agent Based Individual Traffic guidance) system. First I will cover the traffic situation of today. After this follows a section on disruptions. The two layers of ABIT are then described and a section on future perspectives conclude the paper.

The commuting of people is inevitable. Everyday we travel to fulfill demands of our daily life. We work, shop, workout, ferry our children and do many other things that require us to move from one place to another. Usually there is a choice of mode (car, bicycle, public transport) when traveling. In traffic modeling the populace is considered as homo economicus – the rational human. Given a choice the homo economicus will always choose the option providing the most convenience. This can be a very deterministic approach, but through advanced modeling and calibration it is possible to provide valuable insight and forecasts on the traffic of today and tomorrow.

The applications of models to traffic modeling and planning have lead to efficient, necessary and qualified decisions. These same models have repeatedly proven the most basic rule of mass car transport today. In 1952 Wardrop enunciated Wardrop's first principle: *"Under equilibrium conditions traffic arranges itself in a congested network in such a way that no individual trip maker can reduce his path cost by switching routes."* In traffic this is considered to be the Nash equilibrium. In other words, the perceived utility or cost of each route

is the same under congestion. It is important to note that this does not mean that all routes take the same time, only that all routes are perceived equal *to the drivers*. Personal preferences may make one driver choose a more beautiful but longer route while another driver chooses a less time consuming route.

Wardrop's second principle is also interesting: *"Under social equilibrium conditions traffic should be arranged in congested networks in such a way that the average (or total) travel cost is minimized."* As [Ortúzar and Willumsen, 2001] illustrates by a simple example the social equilibrium is 0.5% better in the total travel time than the Nash equilibrium.

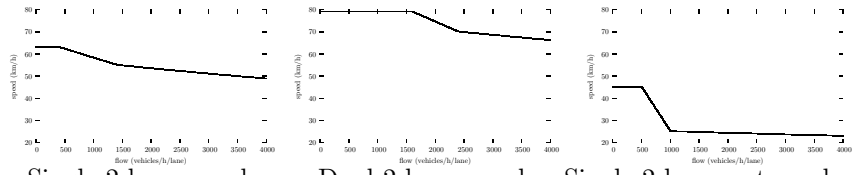
In a recent published article [Roughgarden and Tardos, 2002] on selfish routing the authors find that the Nash equilibrium can be far from the social equilibrium. In fact for complicated speed/flow relationships the ratio between the two equilibria is theoretically unbounded.

A most interesting statement also proposed in the article is: *... to match the performance of a centrally controlled network with selfish routing, simply double the capacity of every edge.*

In most cases drivers choose a route based on the time it takes to traverse it. This is also what most in-car navigation systems and route planning services do. In regular (non-congested) traffic nearly all cars follow the same route from origin to destination. As the traffic flow increases, the speed of a section decreases thus making the preferred route slower. Other routes become attractive and the traffic is diffused into the infrastructure. In this way autonomous vehicle traffic exhibits a form of self balanced sifting.

Interestingly a heterogeneous group of anonymous more or less selfish drivers almost fulfill Wardrop's first principle. Figure A.1 shows traversal times for different types of road section given the flow into the section[1]. As the flow increases so does the traversal time. Here it is worthwhile to examine two

[1]These are generated from the formulas given in [Ortúzar and Willumsen, 2001] page 326.



Single 2 lane, rural Dual 2 lane, rural Single 2 lane, outer urban

Figure A.1: Speed/flow relationship for different road sections

routes at the same time.

Consider two routes, A and B, obeying the single 2 lane outer urban speed/flow relationship. At some point in time route A is 10% faster than B (flow 940 and 1000 respectively). Given the slope on figure A.1, which at the onset of rush-hour is steep, only 3% of the cars (30) shall change route to obtain the Nash equilibrium.

For most networks there is as mentioned above a better social equilibrium, but this cannot be obtained by the action of any individual driver. Some cars must choose a worse route in order to make other cars get a sufficiently better route to achieve the social equilibrium. The situation can be described as a solution space with a local minimum (the Nash equilibrium) and a global minimum (the social equilibrium) which do not coincide. The problem is that the solution method, selfish routing, is incapable of sustaining the global minimum even if the flow distribution should occur. The problem described above is best reflected in the Prisoner's Dilemma, where no single prisoner can ensure himself a better outcome.

Returning to the models and the forecasts, today no country or major city can expect to see a decrease in the number of car-based commuters unless (more or less) drastic countermeasures are applied. This means that the length and intensity of congestion will become increasingly troublesome.

Congestion in it self is self balancing as described above. The problem is that it only balances properly under normal conditions. If the state and usability of the infrastructure diverges from general perception then the self balanced sifting can in worst case be replaced by a very severe jam or even complete traffic grid lock. At this point the driver becomes the most important part in the relative success in the progression of the traffic.

A.1.1 The Driver

As described earlier the mass of drivers tend to distribute themselves according to the Nash equilibrium. This happens primarily because each driver can be seen as an autonomous entity. At some point the driver will try another route and if perceived to be better then it will become the new preferred route. The term "preferred route" is used to reflect what is often encountered in traffic modeling: conservative drivers. Drivers stick to a specific route and if content with it they will seldom try other routes. This means that if the infrastructure changes then it will take some time before the equilibrium is reestablished. Trying a new route may be caused by several reasons, eg. curiosity, impatience, or advice.

The driver may try something new for a change, become unsatisfied with the current route (to slow, long, boring, etc.), or the driver might have heard of a better route.

Above, this conservative attitude is presented almost as having a negative impact on the traffic, but this is not entirely so. If drivers are not conservative, the equilibrium would not become steady as too much of the flow would change path every day.

The behavior of drivers is important to mention. Although conservative attitude is good for the equilibrium, inconsiderate self righteousness is very bad for general flow. The steady state is still obtained by being conservative. The problem is that the often inconsiderate behavior might make the aggressor get a little bit (under congestion a very little bit) faster ahead, but at the expense of everybody else. The result is that the capacity of a given intersection is impeded by selfish or inconsiderate behavior.

The conservative attitude has besides the equilibrium conservation at least one other important impact. Recent research [Abdulhai and Look, 2003] shows that proposing new routes must be done with great care as it is shown to have impact on general safety. New routes mean that people do not know the local traffic conditions and evasion routes (bypassing congestion) are more complex[2].

A.1.2 The Communication

When driving to a predetermined destination we usually have a route or a very small subset of routes in mind; a preferred set of paths to the destination.

As mentioned above this subset of routes tends to be static. When altered, it is usually because we think or know that another route is better. The question is how can it be known if some route is better than another.

This can be done either by guessing, by becoming sufficiently unsatisfied with the current route or be told. Not so many years ago the latter was only performed primarily by the speaker on the traffic radio. This reaches many drivers, but it requires that the driver listen to the radio and it is the right station or the driver has a car radio that can switch to traffic announcements automatically.

Recent advancements in communication and surveillance have made it possible to add new and entirely different forms of communication to the “telling”.

[2]In the article the authors explicitly point out that the number of turning movements in non-preferred intersections is essential to the number of accidents under high load.

Some of these are dynamic message signs (DMS)[3] and real time status messages available over fax, short message service (SMS) or World Wide Web. See [MTO, 1999] or [DRD, 2004] for examples of the latter.

All of this information has increased safety as well as the utilization of the infrastructure.

A.2 The Future for Road Based Traffic

Over the recent years the amount of vehicles concurrently in transit has increased at an alarming rate. In Copenhagen, a capital with public transportation, a recent rapport [HUR, 2004] from The Greater Copenhagen Authority has shown that the average traveling speed on the highways during morning rush hour has dropped 15% to $37km/h$ in one year. As the critical mass of vehicles are approached on the highways so is the traffic in the center of the city. Traffic jams and grid locks are there to stay.

If we focus on the increasing vehicle traffic and leaving out alternatives two options are possible. Either increase the capacity of roads, which is immensely expensive, or change the usage of roads, which is equally complicated.

The latter is already attempted through the previously described communication media and through the use of dynamic intersectional control. The problem is that in most cases the individual driver has to decide solely by himself what to do in different traffic situations.

The idea of ABIT is to address this issue. Instead of forcing every driver to make his own choice unaided the system can propose a set of alternate routes. These can be based on far more information than can be communicated to the driver and thus work as decision support for route choice.

I have previously described how equilibrium was established between possible routes under normal conditions. Intelligent vehicle-wide routing will decrease the time to reach the equilibrium, thus reducing the overall transport cost.

The problem addressed in my PhD study is *what if it is not under "normal conditions"*. In theoretical terms, this means the system is disrupted or suffering from a disruption.

[3] Dynamic message signs are also called variable or changeable messages signs.

A.3 Disruptions

A disruption is a state where the expected best routes are significantly deteriorated.

In other texts on traffic these states are simplified to incidents. Distinct occurrences that are easy to identify and classify. A disruption is a much more general term for disturbances in the system.

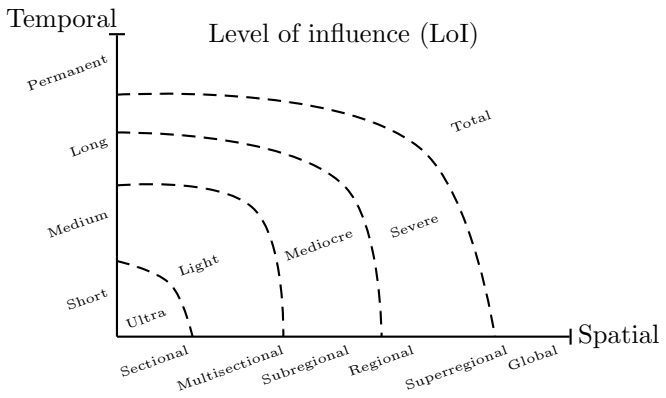
The ABIT system is meant to cross the gap between dynamic traffic guidance and Operations Research (OR). The field within OR that is combined with traffic is *Disruption Management* (DM). DM researches in solution methods for resource efficient replanning when unforeseen events occur. For further information see [Clausen et al., 2001].

Disruptions come in many shapes and sizes and a few of them can be:

- Accidents.
- Emergency vehicles.
- Demonstrations, parades or sport events.
- Congestion, jams and grid locks.
- Road work.
- The weather.

The main problem with disruptions is that they more or less immediately reduce the capacity of some part of the road network. If the load is below the new capacity nothing happens. The infrastructure can still accommodate the demand. The problem will occur if the load is above the new capacity. Depending of the degree of the overloading, queues will build up and the traversal time for routes affected by the queuing will increase, thus making them less attractive. This is exemplified in congestion where the significant presence of vehicles results in performance degradation.

Every disruption can be characterized in two ways. Firstly it can be described by its spatial representation – how large is the geographic area that is affected and if it is moving. Heavy snow is usually regional or global. An emergency vehicle is sectional and roaming as it moves around. A severe accident in an intersection may be multisectional or larger.



The figure shows the mean value for the stochastic LoI function. The stochasticity is a result of the condition of the network and unforeseeable consequences.

Figure A.2: Classification guide

Secondly, it can be described by is temporal locality – the duration of the disruption. Minor accidents which are quickly alleviated exhibit short temporal locality, whereas major construction work is an example of long temporal locality. Similar to roaming spatial locality the term periodical temporal locality is used for repeating disruptions with a fixed schedule.

Even though useful, the two characterizations are not sufficient. A stochastic function on the spatial and temporal locality, the Level of Influence (LoI), can here be introduced. The LoI of a disruption is the impact on the total system. It is generally a measure of how much that must be changed to achieve a new equilibrium. This function is the only measurement that actually assesses the influence on the system and not only the disruption itself. An example could be a sectional and medium accident under light traffic. This could be interpreted as light or ultra light as only local dispersion is needed to achieve a new equilibrium. On the other hand, the same accident during rush hour may exhibit severe LoI as the flow of diverted vehicles has nowhere to go. This usually causes long queues, which in some cases ultimately result in a grid lock.

Figure A.2 is a visualization of the characterizations and the mean value of the LoI. The stochasticity of the function is introduced to cover examples as described above, where the same temporal and spatial locality can lead to different levels of influence.

The problem with a disruption is that it by definition is hard to predict. If

it could be foreseen it would be considered advance knowledge. Knowledge that can be incorporated into the system in proper time and would potentially remove a disruption before it occurred.

A significant issue with a disruption is the driver's reaction to it. Will the driver stay on the current route or will another route be chosen? It is not an easy task to make the driver choose the best route.

An abundance of information must be communicated to and analyzed by the driver if a qualified decision is to be made. This could become a significant safety hazard as cellular communication has proven to be. Here ABIT becomes interesting since it can act as decision support for the individual driver. There is much information available in the system, such as:

- The nature of the disruption
- Traveling speed for all roads in the network
- Optimal undisrupted route
- Optimal route given current conditions
- Impact of disruption over time
- Flow rates in the network over time

This is analyzed and only relevant information is presented to the driver. In the following sections I will cover the main ideas of ABIT in different situations.

A.3.1 Immediate Dispersion

The first and most obvious application is the immediate alleviation of a disrupted situation.

Consider a simple setup as in figure A.3 on the following page. Vehicle A has just encountered a queue on its way to destination D_A. In the setup there are three immediate alternatives: 1) stick to the original route; 2) choose the upper route; or 3) choose the lower route. For some reason the driver chooses the lower route thus running into another queue. What the driver did not know was that the traversal times for each route was 1) 10 minutes; 2) 7 minutes; and 3) 11 minutes. In this case the driver would have saved 4 minutes if information was available.

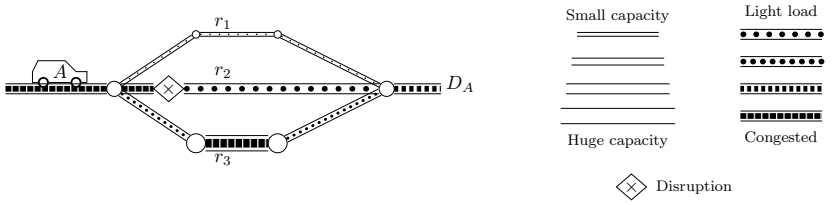


Figure A.3: Example setup with one disruption

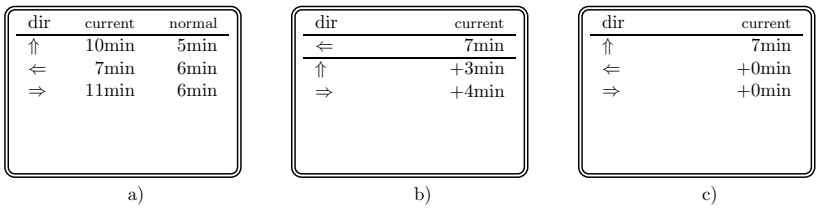
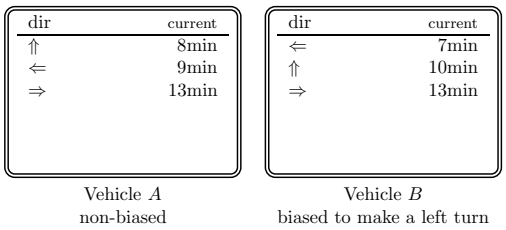


Figure A.4: Examples of dashboard information

The essence of the decision is that it has to be made within a very short time. From A's perspective the decision is immediate, hence the term *immediate dispersion*. A driver in an ABIT equipped vehicle could be presented with the information on the dashboard as it is depicted in figure A.4 a). Thus allowing the driver to easily make a qualified choice. In theory, the system could at every intersection indicate the benefit/loss for choosing a significantly different route as shown in A.4 b), thus implicitly promoting dispersion even when the system is undisrupted. In an non-disrupted congested system the dash should look like A.4 c) if the Nash equilibrium has been obtained.

Dispersion of a single vehicle is rather simple; find the new best route and propose that to the driver. Many people will choose accordingly. Others conclude that if the difference is 3 minutes, they will stick to the direct route. Finally, some may choose the right turn as they do not mind the extra time or simply just like the detour. Inspecting the speed/flow diagrams in figure A.1 on page 79 shows that increasing the flow on a route most likely increases the traversal time of that route. If ABIT directs all vehicles to the upper route it too will become congested. Consequently, as the upper route has less capacity than the other two the impact can be much more severe than the 3 minutes saved in the single case.

Considering A's choice the crucial point is to divert exactly enough cars to make the upper and the direct route become equally time consuming. Given speed/flow graphs for the routes the equilibrium flows can be determined trivially or in not too complex cases it can be done rather quickly by microsimula-



The two vehicles A and B are in the same queue heading for approximately the same destination.

Figure A.5: Example of selective biased dashboard information

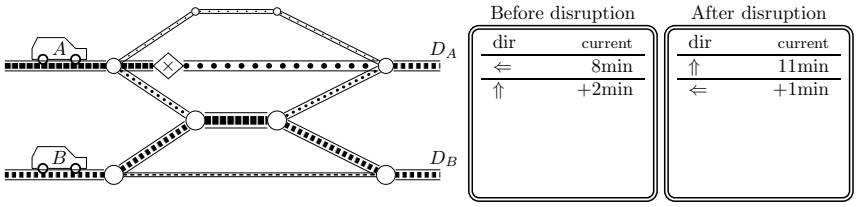


Figure A.6: Extended setup with one disruption and dashboard information for vehicle B

tion. The problem is to make the driver actually choose the right alternative. One approach to achieve the flow split operation could be to trick the drivers. Figure A.5 shows a possibility of this selective biased decision support. This incurs one very significant problem – the credibility of the system. If the drivers do not trust the system then it will be of no use. Selective biasing must thus be used with great care and consideration.

It might be the best strategy to actually inform the driver that there are several different biased propositions, but the exact one is chosen for the current driver to promote a specific route choice.

Up to this point we have only considered immediate single route traffic and left interfering routes and traffic out of the considerations. If we broaden the scope and start considering just a little more of the world, then the situation becomes more complicated. Referring to figure A.6, which is an extended version of figure A.3 on the preceding page, consider the added vehicle B. As the system has all ready assessed some vehicles will be dispersed onto the shared section. This will increase the traversal time on it and consequently make the previously faster route slower than the straight alternative. Without ABIT, B would simply

follow the expected fastest route and thus use an extra minute due to lack of information.

Both examples above only illustrate rather small savings, 3 minutes and 1 minute respectively. However consider the number of vehicles that could be involved. In rush hour thousands of vehicles will be on their way in the infrastructure. [TCC, 1999] states that the road system of Toronto accommodated no less than 5,2 million vehicles each day in 1996 (expected to become 8,1 million by 2021). If ABIT helps just 1% of these in saving 1 minute every day, then $\sim$ 870 man hours would be saved every day. In 2021 the number of hours will be $\sim$ 1350.

I believe that both the fraction of affected vehicles and the number of minutes saved is a pessimistic scenario and that they are in reality significantly higher. As the system is not yet developed or tested, the only indication of the possible impact is the effect of a currently widely deployed approach – DMS. As traffic safety is significantly increased the number of incidents are reduced and thus the overall performance of the infrastructure has risen.

So far, we have only considered the operation of immediate dispersion where ABIT acts as decision support to prevent the build up of back queuing. This is not the only intent of ABIT all though it is very important.

In [Lighthill and Whitham, 1955] the concept of kinematic waves are described in relation to vehicle traffic. From detailed formulations based on flows in floods the authors deduce similar behavior on roads. The essence is that at every point where there is a capacity reduction (e.g. disruption) or sudden surge in load (e.g. at intersections) waves of lower-than-average speed are generated. These consequence waves are then propagated through the traffic at different speeds. Some of the waves actually become shock waves due to the capacity and load of the sections it traverses.

A shock wave is characterized as a situation where the vehicles rather suddenly comes to an almost complete halt and then slowly regains speed. According to the article the only way to make a wave fade is by controlling the speed or flow of vehicles just before the wave. The speed can be controlled by asking the drivers to slow down in advance and the flow can be decreased by diverting traffic before the wave.

Even though the immediate dispersion has been inaugurated the disruption is not necessarily dealt with. The simple extension above showed the effect on nearby vehicles. If the disruption has medium LoI or even higher, the number of vehicles and routes affected increase dramatically. At this point immediate information is not sufficient and both theory and practice become increasingly complex as we turn to individual alleviation.

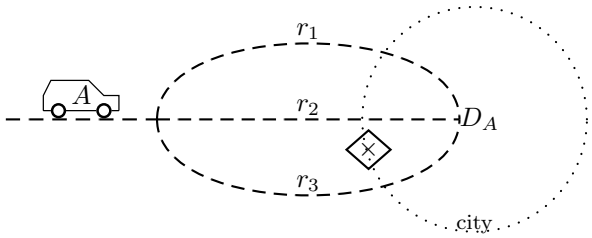


Figure A.7: Individual alleviation example

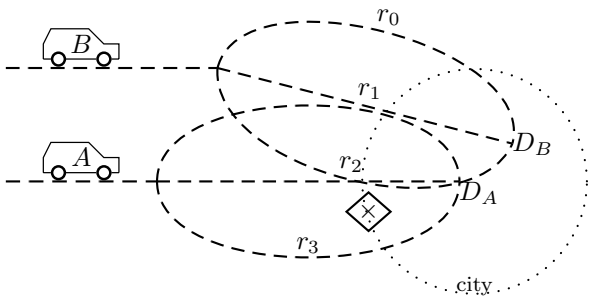


Figure A.8: Extended individual alleviation example

A.3.2 Individual Alleviation

In the previous section immediate dispersion dealt with the local, both spatial and temporal, impact of a disruption. However, individual alleviation is the overall measure to minimize the impact of a disruption. The idea is to divert traffic in a much larger scope than with immediate dispersion.

Consider figure A.7, vehicle A is heading for the destination D_A. As the disruption $\times$ happened earlier it has all ready been assessed to be of severe impact. Dispersion of traffic is thus impeding routes r_2 and r_3. At present route r_1 is best, but as it takes (according to the system) 15 minutes to reach the impeded area the disruption may be alleviated at that time. If that was the case, then both r_2 and r_3 will be faster than r_1. This little scenario depicts a very onerous complication in individual alleviation, the concept of time and propagation. I expect that this combined with the nature of disruption is the crucial success criterion in ABIT. Extensive modeling and forecasting must be applied to get anywhere near a viable approach.

If the disruption is severe it is unlikely that the situation returns to normal within the relatively short time span of 15 minutes. The expected best route for vehicle A is then r_1 in this situation.

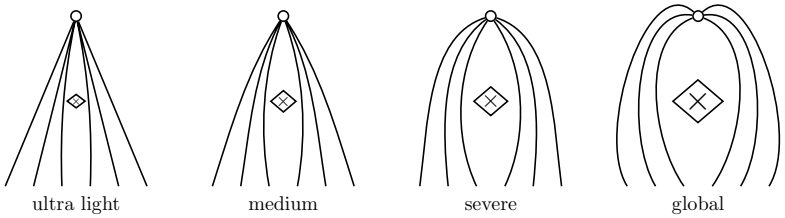


Figure A.9: Level of Influence bending patterns

If we extend the scenario with vehicle B as in figure A.8 on the previous page we again see that simply diverting A might cause secondary disruptions (jams due to insufficient capacity), if B continues onto the direct route. Depending on the prediction of the flow over time for the individual sections of the route, B may arrive sooner by continuing along route r_0.

As with immediate dispersion we bend the routes to accommodate the new flow pattern. Figure A.9 depicts different bending patterns according to a disruption levels of influence.

The problem with redistribution of flow, both immediate dispersion and individual alleviation, is that the disrupted infrastructure might not be able to accommodate the flow. In theory this means infinite queuing, in practice it yields jams and in extreme cases grid locks where the traveling speed is close to zero.

At present, we have only considered dissipating disruptions. It is important to realize the system's potential when experiencing imminent disruptions. An imminent disruption is a disruption that is bound to appear and is known in advance. An example is roadwork. A specific section or intersection has reduced capacity as new tarmac must be applied. Outside rush hour this is no problem as the load is below the reduced capacity. Flow forecasting will indicate that if traffic is distributed as usual it will result in a disruption. The intent of ABIT in this case is to reduce the disruption by diverting traffic even before the problem becomes critical. A side effect of this is that the consequence waves are reduced.

A.3.3 Disruption Avoidance

Imminent disruptions is the first step to actual disruption avoidance, which can be interpreted as a proactive remedy to avoid disruptions. Based on accident prediction models, flow forecasting and capacity assessment it might be possi-

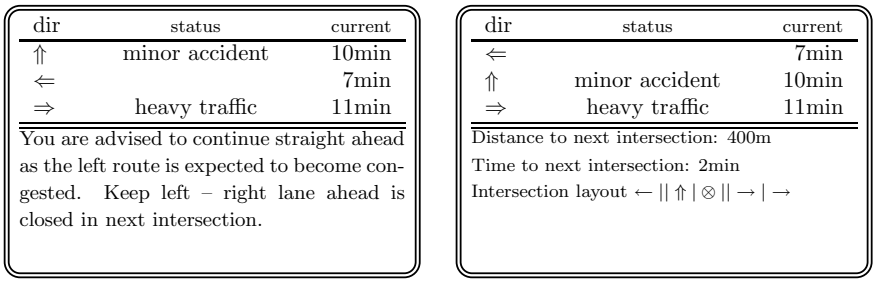


Figure A.10: Examples of enhanced dashboard information

ble to reduce the accident risk, which may be a way to reduce the number of accidents and their severity.

Combined with LIWAS described in [Hansen, 2004] it may prove even more efficient to increase both safety and utilization of the infrastructure. Figure A.10 shows some examples of enhanced information to the driver.

A.4 Further possibilities

Once the system is operational the abundance of information gathered can be put to even further uses. Except for the extreme surveillance possibilities, which are inherent to the system, several other more user-oriented applications are possible.

A very simple yet very important usage is the in-vehicle warning signal on inbound emergency vehicles. In these cases every minute counts and thus appropriate warning of involved vehicles and extended intersectional control may decrease the transit time for the emergency vehicle.

Furthermore, the light controlled intersections can become increasingly intelligent if they are not only programmed to perform a specific schedule. They can also be controlled according to the flow information in the system, thus increasing the vehicular throughput in specific directions to alleviate a disruption or increase overall flow.

The information can also be used for the public transport as road based vehicles can be forecasted with much better precision than today. GIS is getting increas-

ingly implemented in public transport thus allowing for arrival time prediction. However, this is based on aggregate modeling and not on exact information as possible with ABIT.

A more commercially oriented application could be parking lot assignment. The driver selects which part of town or specific lot and the system provides route and reserves a lot for the car.

A.5 Conclusion

Even though the potential in ABIT is encouraging, there are some important issues that must be considered.

Gathering the real time status information is today insufficiently precise, as per section status must be available. [Roughgarden and Tardos, 2002] calculates the impact of imprecise or out-of-date data in the decision process. The conclusion of their research in relation to ABIT is that slightly out-of-date information is sufficient. The system will not need to know the exact state of entire network all the time. It might therefore be sufficient to use only a fraction of the vehicles to gather real time status of the infrastructure.

Immediate dispersion is, when compared to individual alleviation, theoretically simple. In practice the question of the necessary fraction of ABIT-enabled vehicles is crucial. Is it sufficient to have 10% of the vehicles ABIT-enabled to make a difference or is the critical fraction even higher?

The Nash equilibrium example showed that only a fraction of vehicles is necessary to reestablish equilibrium in slightly perturbed situations. I believe that the greater the fraction of ABIT-enabled vehicles the higher LoI of disruptions can be alleviated.

The technical details of communication and server structure as well as in-vehicle implementation is also an unexplored field. Ongoing work on the ex-hoc infrastructure in [Hansen, 2004] might yield valuable insights to this problem.

Given the lack of enforcible incentive (such as the London congestion charge) for the individual driver, I expect that ABIT will only be capable of guiding to the Nash equilibrium and not the social equilibrium. ABIT is inherently depending on the autonomous driver and can therefore only try to make a difference. It is up to the drivers to actually make the difference.

Bibliography

[Abdulhai and Look, 2003] Abdulhai, B. and Look, H. (2003). Impact of dynamic and safety-conscious route guidance on accident risk. *Journal of Transportation Engineering*, 129(4):369–376.

[Clausen et al., 2001] Clausen, J., Hansen, J., Larsen, J., and Larsen, A. (2001). Features - disruption management - how or can get interrupted operations back on track – fast. *OR/MS Today - Operations Research/Management Science*, 28(5):40–43.

[DRD, 2004] DRD (2004). Trim – traffic map [in danish]. Online by Danish Road Directorate. direct link http://www.trafikken.dk/wimpdoc.asp?page=document&objno=77436.

[Hansen, 2004] Hansen, K. M. (2004). The ex hoc infrastructure – enhancing traffic safety through life warning systems. *Trafikdage på Aalborg Universitet*. Available from the conference website http:/www.trafikdage.dk/.

[HUR, 2004] HUR (2004). Regional traffic rapport [in danish]. Technical report, The Greater Copenhagen Authority, www.hur.dk. direct link http://www.ht.dk/86353150-DBE4-450A-BEE4-502096DBB099.

[Lighthill and Whitham, 1955] Lighthill, M. J. and Whitham, G. B. (1955). On kinematic waves. ii. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 229(1178):317–345.

[MTO, 1999] MTO (1999). Traffic and road information system (tris). Online by Ontario Ministry of Transportation. direct link http://www.mto.gov.on.ca/english/traveller/compass/systems/tris.htm.

[Ortúzar and Willumsen, 2001] Ortúzar, J. D. D. and Willumsen, L. G. (2001). *Modelling Transport*. John Wiley & Sons Inc., 3rd edition.

[Roughgarden and Tardos, 2002] Roughgarden, T. and Tardos, E. (2002). Operations research - how bad is selfish routing? *Journal of the ACM - Association for Computing Machinery*, 49(2):236–259.

[TCC, 1999] TCC (1999). Gta city status. Technical report, Toronto City Counsil, http://www.city.toronto.on.ca/. direct link http://www.city.toronto.on.ca/torontoplan/citystatus_5.htm.

Appendix B

Agent Based Individual Traffic guidance – Second Presentation

When working with traffic planning or guidance it is common practice to view the vehicles as a combined mass. From this models are employed to specify the vehicle supply and demand for each region. As the models are complex and the calculations are equally demanding the regions and the detail of the road network is aggregated. As a result the calculations reveal only what the mass of vehicles are doing and not what a single vehicle is doing.

This is the crucial difference to Agent Based Individual Traffic guidance (ABIT). ABIT is based on the fact that information on the destination of each vehicle can be obtained. This information can then be used to provide individual traffic guidance as opposed to the mass information systems of today – dynamic road signs and traffic radio. The goal is to achieve better usage of the road and time.

The main topic of this paper is the current development in both practical and theoretical fields concerning the realization of ABIT.

B.1 Overview

The objective of this document is to present an overview of highly advanced traffic guidance today and especially what remains until Agent Based Individual Traffic guidance (ABIT) is possible. The essentials of ABIT is briefly covered. Following this the most essential theoretical research is covered. This is supplemented by a section on the practical evolution of the hardware required for ABIT. A concluding section will briefly give my expectations on the future of Pervasive Traffic Intelligence (PTI) in general and ABIT specifically.

B.2 Agent Based Individual Traffic guidance

The applications of models to traffic modeling and planning have lead to efficient, necessary and qualified decisions. These same models have repeatedly proven the most basic rule of mass car transport today. In 1952 Wardrop [Wardrop, 1952] enunciated Wardrop's first principle: *"Under equilibrium conditions traffic arranges itself in a congested network in such a way that no individual trip maker can reduce his path cost by switching routes."* In traffic this is considered to be the Nash equilibrium. In other words, the perceived utility or cost of each route is the same under congestion. It is important to note that this does not mean that all routes take the same time, only that all routes are perceived equal *to the drivers*. Personal preferences may make one driver choose a more beautiful but longer route while another driver chooses a less time consuming route.

Wardrop's second principle is also interesting: *"Under social equilibrium conditions traffic should be arranged in congested networks in such a way that the average (or total) travel cost is minimized."* As [Ortúzar and Willumsen, 2001] illustrates by a simple example the social equilibrium can be 0.5% better in the total travel time than the Nash equilibrium.

In the article [Roughgarden and Tardos, 2002] on selfish routing the authors find that the Nash equilibrium can be far from the social equilibrium. In fact for complicated speed/flow relationships the ratio between the two equilibria is theoretically unbounded.

A most interesting statement also proposed in the article is: *... to match the performance of a centrally controlled network with selfish routing, simply double the capacity of every edge.*

In most cases the drivers choose a route based on the time it takes to traverse it. This is also what most in-car navigation systems and route planning services do. In regular (non-congested) traffic nearly all cars follow the same route from origin to destination. As the traffic flow increases, the speed of a section decreases thus making the preferred route slower. Other routes become attractive and the traffic is diffused into the infrastructure. In this way autonomous vehicle traffic exhibits a form of self balanced sifting.

As exemplified in the previous paper [Wanscher, 2004] on ABIT only redirection of a minor fraction of the vehicles is necessary to obtain the Nash equilibrium.

For most networks there is as mentioned above a better social equilibrium, but this cannot be obtained by the action of any individual driver. To achieve the social equilibrium some cars must choose a worse route in order to make other cars get a sufficiently better route. The situation can be described as a solution space with a local minimum (the Nash equilibrium) and a global minimum (the social equilibrium) which do not coincide. The problem is that the solution method, selfish routing, is incapable of sustaining the global minimum even if the flow distribution should occur. This is one of the mayor benefits of introducing road pricing ([Yang and Huang, 2004]), as the pricing might be calibrated to ensure that the social equilibrium is sustainable.

Returning to the models and the forecasts, today no country or major city can expect to see a decrease in the number of car-based commuters unless drastic countermeasures are applied. This means that the length and intensity of congestion will become increasingly troublesome.

Congestion in it self is self balancing as described above. The problem is that it only balances properly under normal or expected conditions. If the state and usability of the infrastructure diverges from general perception then the self balanced sifting can in worst case be replaced by a very severe jam or even complete traffic grid lock. At this point the driver becomes the most important part in the relative success in the progression of the traffic.

B.2.1 The Driver

As described earlier the mass of drivers tend to distribute themselves according to the Nash equilibrium. This happens primarily because each driver can be seen as an autonomous entity. At some point the driver will try another route and if perceived to be better then it will become the new preferred route. The term “preferred route” is used to reflect what is often encountered in traffic modeling: conservative drivers. Drivers stick to a specific route and if content with it they

will seldom try other routes. This means that if the infrastructure changes then it will take some time before the equilibrium is reestablished. Trying a new route may be caused by several reasons, eg. curiosity, impatience, or advice. The driver may try something new for a change, become unsatisfied with the current route (to slow, long, boring, etc.), or the driver might have heard of a better route.

Above, this conservative attitude is presented almost as having a negative impact on the traffic, but this is not entirely so. If drivers are not conservative, the equilibrium would not become steady as too much of the flow would change path every day.

The behavior of drivers is important to mention. Although conservative attitude is good for the equilibrium, inconsiderate self righteousness is very bad for general flow. The steady state is still obtained by being conservative. The problem is that the often inconsiderate behavior might make the aggressor get a little bit (under congestion a very little bit) faster ahead, but at the expense of everybody else. The result is that the capacity of a given intersection is impeded by selfish or inconsiderate behavior.

The conservative attitude has besides the equilibrium conservation at least one other important impact. Recent research [Abdulhai and Look, 2003] shows that proposing new routes must be done with great care as it is shown to have impact on general safety. Traversing new routes mean that people do not know the local traffic conditions. Furthermore routes bypassing congestion are more complex[1].

B.2.2 The Communication

When driving to a predetermined destination we usually have a route or a very small subset of routes in mind; a preferred set of paths to the destination.

As mentioned above this subset of routes tends to be static. When altered, it is usually because we think or know that another route is better. The question is how can it be known if some route is better than another.

This can be done either by guessing, by becoming sufficiently unsatisfied with the current route or be told. Not so many years ago the latter was only performed primarily by the speaker on the traffic radio. This reaches many drivers, but it requires that the driver listen to the radio and it is the right station or the driver has a car radio that can switch to traffic announcements automatically.

[1] In the article the authors explicitly point out that the number of turning movements in non-preferred intersections is essential to the number of accidents under high load.

Recent advancements in communication and surveillance have made it possible to add new and entirely different forms of communication to the "telling". Some of these are dynamic message signs (DMS)[2] and real time status messages available over fax, short message service (SMS) or World Wide Web. See [MTO, 1999] or [DRD, 2004] for examples of the latter.

All of this information has increased safety as well as the utilization of the infrastructure.

B.3 The Future for Road Based Traffic

Over the recent years the amount of vehicles concurrently in transit has increased at an alarming rate. In Copenhagen, a capital with public transportation, a report [HUR, 2004] from The Greater Copenhagen Authority has shown that the average traveling speed on the highways during morning rush hour has dropped 15% to $37km/h$ in one year. As the critical mass of vehicles are approached on the highways so is the traffic in the center of the city. Traffic jams and grid locks are there to stay.

If we focus on the increasing vehicle traffic and leaving out alternatives two options are possible. Either increase the capacity of roads, which is immensely expensive, or change the usage of roads, which is equally complicated.

The latter is already attempted through the previously described communication media and through the use of dynamic intersectional control. The problem is that in most cases the individual driver has to decide solely by himself what to do in different traffic situations.

The idea of ABIT is to address this issue. Instead of forcing every driver to make his own choice unaided the system can propose a set of alternate routes. These propositions can be based on far more information than can be communicated to the driver and thus work as decision support for route choice.

I have previously described how equilibrium was established between possible routes under normal conditions. Intelligent vehicle-wide routing will decrease the time to reach the equilibrium, thus reducing the overall transport cost.

The problem addressed in my PhD study is *what if it is not under "normal conditions"*. In theoretical terms, this means the system is disrupted or suffering from a disruption.

[2]Dynamic message signs are also called variable or changeable messages signs.

The previous submission to Aalborg Trafikdage [Wanscher, 2004] contains a thorough description of disruptions and how ABIT is intended to function.

The remainder of this paper will concentrate on the requirements for ABIT:

- a theoretical base appropriate for understanding, solving and forecasting the mathematical representation of routing vehicles
- software and hardware to support the theoretical methodology
- In-vehicle information systems with communication and visualization capabilities
- Information gathering hardware and software for obtaining and processing real time information from the infrastructure

B.4 Theoretical Requirements

To address a real world problem scientifically usually requires a model. The model then requires a solution method which is equally important. Having a perfect model with absolutely no chance of solving it is just as bad as having a very poor model which is instantly solvable. The key issue is that the model is adequately aggregated to allow both proper representation of reality and efficient solution.

The most basic approach to a model for traffic networks was the Traffic Assignment Problem (TAP). The TAP is from a modeling perspective fairly simple. The crucial difficulty in solving an instance is the selection of a cost function.

A cost function is a function that yields the cost of traversing a link. The cost function can include both time and space to yield a very complicated, but hopefully more representative, function. It is important that we consider how representative a given cost function is and not how precise. If the function is representative of reality the resulting flows will also represent valid, feasible and optimal solutions dependent on the model. However the solution value will only be transferable if the precision is high. Evidently a more complicated cost function is more representative and precise. The difficulty however is that even rather simplistic cost functions require complicated solution methods. Examples of older cost functions can be found in [Ortúzar and Willumsen, 2001] and an example of a newer and significantly more complicated one can be found in [Meschini et al.,].

In the TAP the traditional cost function is both separable and time insensitive. The separability means that the cost of traversing any link is independent of the cost on any other link. The time insensitivity means that the cost of traversing a link is independent of at which point in the considered interval that the link is entered.

These two presumptions seems unrealistic given that rush hour is an excellent example of a situation where entering a link 5 minutes later may result in another 5 minutes of traveling time. The inseparability is easily argued as non representative. The first and most common example is the regular behavior when attempting a yielding turn. That is a turn where the driver has to yield for traffic on other sections. This indicates that the cost of traversing the link that the driver is turning from is dependent on the flow on all the right-of-way links in the given situation. Thus in reality the cost function appears inseparable.

However it still seems that these cost functions are widely used. The evolution from the 1950'ies is that with the increased gathered information from the infrastructure the statisticians are able to calibrate the functions to achieve high precision or representivity in specific cases.

The reason for still using this approach is that the computational requirements are insignificant compared to time sensitive cost functions.

Adding time sensitivity leads to the Dynamic Traffic Assignment (DTA). The DTA is actually still using the same cost function, but the surrounding model is expanded to include time as a constructing part of the model. However, the impact is devastating as increasing the time precision affects the number of variables exponentially and cumbersome algorithms must be used to yield usable results.

Dealing with the inseparability however has been a far greater challenge. Even though mathematicians proved the quasi newton search method many years ago only recent papers are applying it.

The delay from mathematical progress until its application in traffic science is up to 20 years. This is far too long to simply have been unnoticed. My assumption is that even though the theories were developed the actual hardware and software for realizing the theory on large and complicated networks was unavailable.

Through the later years computing power and communication possibilities have increased dramatically. This along with general increase in the implementation skills among researches has resulted in new and interesting approaches to precise and representative traffic modeling.

Examples of different models like Complementarity Problems found in [Isac, 1992], Variational Inequality Problems in [Friesz et al., 1989] or [Patriksson, 2004], hydrodynamic modelling in [Lighthill and Whitham, 1955] or Mathematical Programming in [Meschini et al.,], [Pang et al., 1999], [Ran and Boyce, 1996], [Ran et al., 2002b] and [Psaraftis, 1995] show that a wide variety of approaches are being used. [Peeta and Ziliaskopoulos, 2001] provide a more detailed overview of state of the art within DTA.

The most recent and promising research is [Bierlaire and Crittin, 2006], which applies efficient mathematics to get a solution to inseparable time sensitive traffic models. They also cover the newer variant within DTA, the Consistent Anticipatory Route Guidance.

Solving a single instance is however not sufficient to actually support ABIT. It will be sufficient for the immediate dispersion, which only concern a limited area for a short duration, but two questions remain.

Every model constructed considers a specific period of a daily or weekly cycle. As ABIT is pervasive and continuous a definition of the beginning and the end is not directly obtainable. This however is readily addressed by statisticians. Statistical modeling may be used to find critical periods within which a traffic model cannot "end". Extending the period considered beyond this interval will thus allow us to use the regular Traffic Assignment models at an increased time interval. As mentioned above time can have a devastating impact on the solution time. However for immediate dispersion the networks considered are small and the effect is limited.

The research in [Ran et al., 2002a] is also interesting as they consider the issue of non-completed trips.

On the other hand individual alleviation requires consideration of large networks and forecasting. Forecasting is complicated, but as the information is gathered intrinsically in ABIT model construction and calibration will be improved.

Another troublesome issue in individual alleviation is the size of the considered network. However as forecasting introduces some uncertainty actually knowing a section by section path far away is irrelevant. The key issue is an estimate on the time passing through different areas. An overall route is thus planned and as the driver travels along the path all the exact path is known for the local area. The rest of the network is reduced to a number of aggregated junctions that will reduce the number of links and consequently allow us again to use the presently developed models and solvers.

B.5 Computing Strength

The computing strength has greatly increased over the recent decades. Almost every resource available to the programmer has been increased. At present the computing power is sufficient for solving immediate dispersion and possibly coarse arealizations in individual alleviation. However as the computing power will continue to increase both the immediate dispersion and the individual alleviation will become more representative.

Considering grid computing is possible, but given the obviously very short response times it seems that dedicated computing is preferable. Grid computing may still be applied to pre-calculate the effect of the most likely disruptions and thus reduce the load on the dedicated servers in those cases.

The increase in both wire-based and wireless communication bandwidth will allow for both distributed and parallel approaches.

B.6 Vehicle Information Systems

The field of vehicle information systems has evolved significantly over the recent years. From the first high tech version simply telling where you are to the more sophisticated versions today that can propose alternate routes and also illustrate intersection layout.

Most of these information systems however are utilizing only static information of the network. This is as described above insufficient in most larger cities where disruptions occur on a daily basis. Whether the availability of car navigation results in greater frustration when it gives wrong advice is beyond this text. Lately Traffic Message Channel (TMC) as described in [Forum, 2004] is an attempt to alleviate exactly this by automating the distribution of information concerning disruptions.

The existence and increasing usage of TMC enabled devices shows that the technology to distribute information is already deployed and functional. However the delay from detection to broadcast in TMC is significant and coarse compared to the real time possibilities in ABIT.

The newer handheld navigation devices are also promising as they provide not only increased computing power, but also two way communication from the handheld to the Internet by either Global Packet Radio Service (GPRS) or

Wireless Local Area Network (WLAN). These technologies however are either costly or not yet deployed sufficiently to actually support ABIT. With the programming tools today it will be possible to make a system that seamlessly can change between any available wireless communication.

Advanced compression is standard in almost any operating system and algorithms for minimizing the amount of transmitted information is available. Along with multicasting this may lead to both efficient and reliable communication.

Freescale's MobileGT Total5200 hardware unit [Freescale, 2004] for vehicles is an example of the level of possible integration.

B.7 Information gathering and processing

The information is one of the most crucial parts of ABIT. Poor information results in poor representation in the model and thus inadequate flow distributions. This in turn will result in distrust in ABIT. At present ABIT does not include any sanctioning possibilities and distrust will therefore result in ignorance, in which case the system is ineffective.

Gathering the information is readily done manually by video surveillance or semi automatically by sensors imbeded in the road surface. Advanced statistics are applied to this data to increase the usability of the obtained information.

Every single unit in ABIT is capable of obtaining and sending real time information. The coverage and detail of infrastructure status will greatly increase as the number of ABIT enabled vehicles increases. Gathering the informations is thus intrinsically solved by ABIT. The question here is how to handle the possibly devastating mass of information intelligently. However, this is already considered in the hierarchal construction of the solver. Only the local information is detailed. Information on more distant area is aggregated and the information handled at any single area is thus primarily the local information, which naturally is relatively limited. [Roughgarden and Tardos, 2002] calculates the impact of imprecise or out-of-date data in the decision process. The conclusion of their research in relation to ABIT is that slightly out-of-date information is sufficient. However great care must be taken as [Arnott et al., 1991] concludes faulty information leads to exacerbation of congestion in some areas.

B.8 Conclusion

Since the beginning of the ABIT project three years ago the number of relevant publications has increased dramatically. Movement in the industry as well as the increased rush hour impact for every mayor city indicates that ABIT-like systems are not far from here.

In practice the question of the necessary fraction of ABIT-enabled vehicles is crucial. Is it sufficient to have 10% of the vehicles ABIT-enabled to make a difference or is the critical fraction even higher?

This point is crucial in the future development of real PTI. Having 10 different systems run by different operators requires significantly more enabled vehicles. Hopefully governmental authorities will realize the necessity of highly advanced and cooperating PTI and enforce a common standard as inaugurated with TMC.

Given the lack of enforcible incentive (such as the London congestion charge) for the individual driver, I expect that ABIT or any PTI will only be capable of reestablishing the Nash equilibrium and not obtaining the social equilibrium. ABIT is inherently depending on the autonomous driver and can therefore only try to make a difference.

Bibliography

[Abdulhai and Look, 2003] Abdulhai, B. and Look, H. (2003). Impact of dynamic and safety-conscious route guidance on accident risk. *Journal of Transportation Engineering*, 129(4):369–376.

[Arnott et al., 1991] Arnott, R., De Palma, A., and Lindsey, R. (1991). Does providing information to drivers reduce traffic congestion? *Transportation Research, Part A (General)*, 25A(5):309–318.

[Bierlaire and Crittin, 2006] Bierlaire, M. and Crittin, F. (2006). Solving noisy, large-scale fixed-point problems and systems of nonlinear equations. *Transportation Science*, 40(1):44–63.

[DRD, 2004] DRD (2004). Trim – traffic map [in danish]. Online by Danish Road Directorate. direct link http://www.trafikken.dk/wimpdoc.asp?page=document&objno=77436.

[Forum, 2004] Forum, T. (2004). Tmcforum.com. Online by TMC Forum. direct link http://www.tmcforum.com/.

[Freescale, 2004] Freescale (2004). Freescale mobilegt total5200 product brief. Freescale. direct link http://www.freescale.com/files/microcontrollers/doc/prod_brief/MOBILEGT5200PB.pdf.

[Friesz et al., 1989] Friesz, T. L., Luque, J., Tobin, R. L., and Wie, B.-W. (1989). Dynamic network traffic assignment considered as a continuous time optimal control problem. *Operations Research*, 37(6):893–901.

[HUR, 2004] HUR (2004). Regional traffic rapport [in danish]. Technical report, The Greater Copenhagen Authority, www.hur.dk. direct link http://www.ht.dk/86353150-DBE4-450A-BEE4-502096DBB099.

[Isac, 1992] Isac, G. (1992). *Lecture notes in Mathematics: Complementarity Problems*. Springer-Verlag.

[Lighthill and Whitham, 1955] Lighthill, M. J. and Whitham, G. B. (1955). On kinematic waves. ii. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 229(1178):317–345.

[Meschini et al.,] Meschini, L., Bellei, G., Gentile, G., and Papola, N. An implicit path enumeration model and algorithm for dynamic traffic assignement with congestion spillback. Technical report, Dipartimento di Idraulica Trasporti e Strade, University of Rome 'La Sapienza' Italy. Publish year unknown.

[MTO, 1999] MTO (1999). Traffic and road information system (tris). Online by Ontario Ministry of Transportation. direct link http://www.mto.gov.on.ca/english/traveller/compass/systems/tris.htm.

[Ortúzar and Willumsen, 2001] Ortúzar, J. D. D. and Willumsen, L. G. (2001). *Modelling Transport*. John Wiley & Sons Inc., 3rd edition.

[Pang et al., 1999] Pang, G., Takabashi, K., Yokota, T., and Takenaga, H. (1999). Adaptive route selection for dynamic route guidance system based on fuzzy-neural approaches. *Vehicular Technology, IEEE Transactions on*, 48(6):2028 –2041.

[Patriksson, 2004] Patriksson, M. (2004). Sensitivity analysis of traffic equilibria. *Transportation Science*, 38(3):258–281.

[Peeta and Ziliaskopoulos, 2001] Peeta, S. and Ziliaskopoulos, A. K. (2001). Foundations of dynamic traffic assignment: The past, the present and the future. *Networks and Spatial Economics*, 1(4):233–265.

[Psaraftis, 1995] Psaraftis, H. (1995). Dynamic vehicle routing: status and prospects. *Annals of Operations Research*, 61:143–164.

[Ran and Boyce, 1996] Ran, B. and Boyce, D. (1996). *Modeling dynamic transportation networks*. Springer, Heidelberg, Germany.

[Ran et al., 2002a] Ran, B., Lee, D.-H., and Shin, M. S.-I. (2002a). Dynamic traffic assignment with rolling horizon implementation. *Journal of Transportation Engineering*, 128(4):314–322.

[Ran et al., 2002b] Ran, B., Lee, D.-H., and Shin, M. S.-I. (2002b). New algorithm for a multiclass dynamic traffic assignment model. *Journal of Transportation Engineering*, 128(4):323–335.

[Roughgarden and Tardos, 2002] Roughgarden, T. and Tardos, E. (2002). Operations research - how bad is selfish routing? *Journal of the ACM - Association for Computing Machinery*, 49(2):236–259.

[Wanscher, 2004] Wanscher, J. B. (2004). Agent based individual trafficguidance. In *Trafikdage 2004 (http://www.trafikdage.dk/)*.

[Wardrop, 1952] Wardrop, J. (1952). Some theoretical aspects of road traffic research. *Institution of Civil Engineers – Proceedings*, 1:325–362.

[Yang and Huang, 2004] Yang, H. and Huang, H.-J. (2004). The multi-class, multi-criteria traffic network equilibrium and systems optimum problem. *Transportation Research, Part B (Methodological)*, 38B(1):1–15.

Appendix C

ABIT Measuring

Abstract

This working paper is a discussion on the possibilities for representing reality in a model through different measurable characteristics, aggregation and objective functions.
The measurable characterstics are viewed in relation to type and usage of the information that they yield.
Aggregation considers the process of reducing the information gathered. Evidently some detail is lost and we cover some approaches and their benefits and drawbacks.
The objective function is crucial to any Operations Research (OR) model as it defines the solution value of any feasible point. The direction or sense of the values in the solution space is also defined through the objectives function. We consider several alternatives and discuss them in relation to Agent Based Individual Traffic guidance (ABIT)

C.1 Introduction

This paper is a discussion on different modelling approaches from an OR perspective.

To choose between one or more alternatives among others is not necessarily complicated. The troublesome step is the argumentation for the decision. It is not always an option to just claim some alternative to be superior to the others.

In the research on ABIT, algorithms will be considered and a goal is to select the best of these. To do this a proper measurement framework must be available. With out the framework any decision, even a correct one, will be an unsubstantiated assertion.

We thus have to decide on the framework. This has to be done with great care and consideration as the construction of the framework will have decisive impact on the outcome of the research. This paper concentrates on defining the objective of the ABIT system.

As mentioned above a framework introduces comparability and substance to a decision. This is central to all problems in science. Once we have decided on the framework the decision can be made by deciding the sense of the problem. All we have to do is ask for the smallest, cheapest or most durable solution, and the framework will result in a ranking of the solutions from which we can choose[1].

The problem with any measure and measuring framework is that the output must be interpreted and used correctly. Lack of insight in the framework can lead to nonsensical interpretations and erroneous decisions. This paper not only represents the objective of ABIT it also sheds light on the constituents of the framework and their interaction.

Whenever we consider a measure we have to consider the point of view we take. Are we considering this from an overall perspective, or are we considering the individual that experiences the impact of the decision based on the framework. It is not hard to make one vehicle travel fast – the problem is that all other drivers will be significantly impeded by that decision. This distinction is essential in the construction of the objective, which thus will have elements of multi criteria optimization.

This paper consists of three main sections on measurable characteristics, aggregation and objectives. Each of these covers areas in the framework. First the atomic parts of the framework – the measurable characteristics – are covered. Then the section on aggregation describes approaches to combine or process the excessive amount of information acquired from the measured data. The section on objectives concludes the construction by combining the aggregations

[1] This is the case in single objective optimization. In multi objective optimization the ranking can be a far more complex ordering of the alternatives.

and measured data.

C.2 Measurable Characteristics

The concept of a measure is simply to have a value on an item or situation based on some criterion. If we consider the most obvious for traffic, they are:

Velocity The velocity reflects the traveling speed of the vehicle. For immediate situational evaluation it is sufficient. The problem is that it does not consider increased traveling distance or fluctuations in the velocity over time.

Travel time This is the duration of the trip or subpart of the trip. Even though it is good for general views of the network, it cannot be used for immediate evaluation, as the measure reflects a history and not the immediate condition. Furthermore the travel time measure has a serious flaw which is described in [Roughgarden and Tardos, 2002]. The effect of simplistic travel time evaluation is actually a significant increase in the total travel time of all vehicles.

Flow This measure represents the number of vehicles passing a specific area or point per time unit. Dependent on the time unit the measure can be both immediate and history based. The problem is that is does not tell anything on any specific vehicle. The same flow can be obtained by low speed high density traffic and high speed low density traffic.

Density This is an indication of how condensed the traffic is. The higher the density, the more vehicles per meter of road. This measure can only tell immediately if we are experiencing heavy or light traffic. It does not tell if the traffic is moving or stopped.

Distance of travel is the distance moved to get from origin to destination. As with travel time we can only get a historical view and not an immediate reading on the network. Despite the physical relevance of moving as little as possible, it yields no information on the general performance of the network. Sending all vehicles on the shortest path will surely result in insufficient use of all sections not part of a shortest route. Particularly it may result in catastrophic problems as the trunk roads are not always the shortest.

These are, although descriptive, not all the measures that can be relevant. If we turn our attention away from the direct physical measures, we can consider:

Quality conditions can be divided into two. *Obnoxious conditions* covers incidents that have a negative impact on the trip experience. This could be queuing, complicated routes, excessive number of left turns or weather conditions. Individuals may defer highways due to discomfort at high speeds. And *Preferred conditions* are the opposite of the above. This does not include the already covered direct physical measures. It includes road quality, air quality, and visual splendor, etc.

Accident avoidance ([Abdulhai and Look, 2003]) is a fairly new consideration in route guidance. Basically the idea is to include accident prediction models (APMs) in the route planning and attempt to minimize the theoretical risk of accidents.

Environmental concerns includes the indirect impacts of traffic such as pollution and noise.

Political incentives are decisions that can be used as a guide or may be directly contradictory to other measures. An example can be to keep the traffic on the trunk roads even though the distance measure may indicate that this is a suboptimal alternative. Diverting traffic from school-roads is also considered a political incentive.

In all the cases above we have mostly considered the controllers view of the network and not the drivers perspective. The question is how does the driver perceive the trip alternatives.

In most traffic modeling today the route assignment is addressed through advanced discrete choice models ([Ortúzar and Willumsen, 2001]). The idea is to create a model of the perceived utility of each alternative to each individual.

In our case the alternatives are different routes and the individuals are the drivers of the vehicles. As an example, a linear utility function yields the utility U of alternative i to individual j based on k measures or measurable prerequisites[2], where x_{jk} is the taste-parameter for j concerning measure k. Θ_{ik} is the value of the k^{th} measure for alternative i:

$$U_{ijk} = \sum_k x_{jk}\Theta_{ik}$$

In slightly other terms the linear utility function is an individually weighted combination of the measures. It is then expected that the individual will choose the alternative with the highest utility.

[2]This can be network independent socioeconomic characteristic like household income, household size or marital status

It would be most interesting to see if the ABIT algorithms can handle individual preferences and actually use them beneficially.

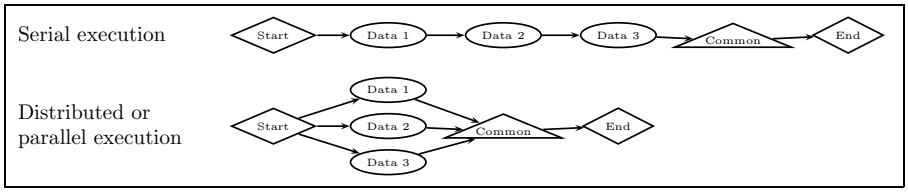
At this point we use previous route selection surveys to construct an expected driver to represent the populace. Many papers on applied Discrete Choice Modeling generally depict time and distance measures as most significant for route choice. The most significant non-economic measure was in [Bovy and Stern, 1990] found to be scenery.

What the papers also indicate is that the models are non-transferable. Even though time is most significant in both Copenhagen and Berlin the taste-parameter may vary greatly. Hence, we cannot make a generalized driver for all locations or individuals.

All significant measures are relevant for a fully informed decision. The problem is that considering all the information is intractable to even the most advanced computers today. The next section on aggregation will describe some approaches to reducing the amount of raw data and the consequences of the reduction.

C.3 Aggregation

The basic idea in aggregation is to reduce the amount of data in consideration. Instead of considering 1 megabyte of data, we can aggregate the information to under 8 bytes. Obviously considering 8 bytes requires less computing power. We still have to reduce the information from 1 megabyte to 8 bytes, but that process is detached from the main calculation. This means that it is both easily parallelizable for different data sources and distributable. Characteristics that are important in any pervasive system, such as ABIT.



Aggregation is not purely beneficial. The reduction in amount of data looses information. If the reduction of data is lossless, it is called compression. This difference is clear when considering compressed data and aggregated data. The aggregated data is usually easy to understand and can be interpreted immediately. The compressed data requires complete decompression to reveal it's

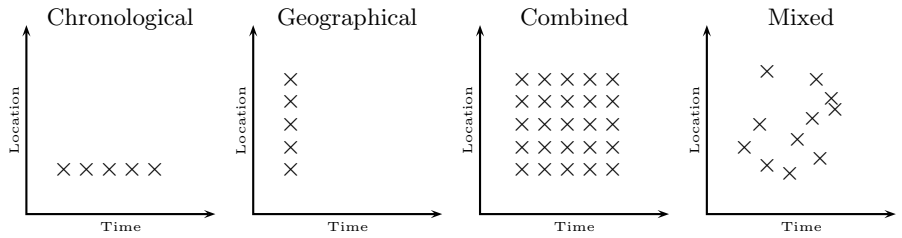


Figure C.1: Generic methods for obtaining measures

secrets.

The crux in aggregation is to balance the loss of information against the reduction of data volume.

For the remainder of this section we consider a set M of n similar measures m_i. The set of measures can be obtained in several ways:

Chronological They all are collected over time from the same location.

Geographical They all are collected at the same time from different locations. This can be considered a snapshot of the system.

Combined The set consists of several snapshots.

Mixed Data are collected from several locations at different times.

The chronological approach yields the history of a single location, whereas the geographical shows the immediate state of the entire set of locations. The combination of these, the combined approach, gives the history of system states. Mixed sets consists of measures that do not exhibit either chronological nor geographical properties. Figure C.1 depicts the classification.

Once the data are obtained there are several ways to reduce them. The most general approaches are by these transformations:

Minimum The set of measures is represented by the smallest element:

$$m_{min} : \{m_{min} \in M, m_{min} \leq m_i, \forall i\}$$

Maximum The set of measures is represented by the largest element:

$$m_{max} : \{m_{max} \in M, m_{max} \geq m_i, \forall i\}$$

Average The set of measures is represented by a measure not necessarily in M with least total linear distance to all other measures.

$$m_{avg} = \frac{\sum_{i=0}^{n} m_i}{n}$$

Sum The set of measures is represented by the sum of all elements.

$$m_{sum} = \sum_{i=0}^{n} m_i$$

All these reduce the amount of data significantly, but they also loose a lot of information. None of the above or a combination of them yields anything about the distribution or correlation. In statistics these are often applied:

Population Variance which describes the spread of the measure points around an average.

Population Standard Deviation which is the square root of the population variance.

Correlation signifies the similarity or difference between two sequences.

All of the above do in general provide a snapshot of the situation aggregating a series of value to a single value. Of course it is possible to make subset aggregation, whereby the entire set of values are not represented by a single value, but by a new and hopefully smaller set of values. However this does not yield information on the trends of the measures. The usual approach to this is to visualize the values in graphs or diagrams. While visualization helps us perceive the values and deduce information, a computer is unable to "get ideas" in the same way.

Instead the computer's trend search must be based on theory. This is where curve fitting becomes interesting.

Curve fitting is the representation of the set of values by a curve. The simplest is linear curve fitting which in d dimensions require only $2d$ values – a d-dimensional offset vector and a d-dimensional slope vector. This can be extended to plane fitting or more complex constructions like a bezier curve.

Curve fitting is however not especially interesting here. It can be used for extrapolation and forecasting, but the computational overhead is significant. Using the curve parameters in calculations in stead of the projection values is

complicated. Nevertheless in the future it might be possible to use this beneficially when solving complex problems with immense amounts of data.

Comparing and combining measures is important. When working with other than chronological data an obstacle is to ensure comparability in the set of measures.

Case	Vehicles/h		Speed		Average speed	
	Highway	Trunk road	Highway	Trunk road	Exact	Relative
Normal flow	5000	2500	110	60	93	100%
Highway impeded	3500	2500	90	60	62	71%
Trunk road impeded	5000	1500	110	45	95	94%
Both impeded	3500	1500	90	45	77	80%

The table shows four cases for the condition of two roads. The first column describes the case. The second how many vehicles that cross the section per hour. The third shows the speed of the vehicles. The fourth column shows on the left the exact average speed of a vehicle on either section whereas the right side shows the average relative speed.
The table shows two different undesired effects. First, the exact average speed is increased when the trunk road is impeded. And second, the average relative speed is higher if both are impeded and not just the highway.

Table C.1: Relative representation

It makes no sense to average mindlessly over two sections with significantly different characteristics. Table C.1 exemplifies the pitfall and a relatively simple attempt at a solution – the locally relative representation. In this every measure is substituted by its position on a scale according to a predetermined scale definition. An example could be to use the maximum aggregate to define the scale, i.e.

$$m_{i_{rel}} = \frac{m_i}{m_{max}}$$

It is important that any relative transformation are performed properly.

The major issue in table C.1 is that the measures used (speed and flow) are not independent. Using instead lane density, the number of vehicles per kilometer per lane, and the speed yields another picture shown in table C.2 on the facing page.

The two simple tables show that great care must be taken to actually get a usable aggregated measure. This is important both when combining several different measures and when aggregating over identical measures.

Having covered both the measures and some of the available aggregation meth-

Case	Lane density (veh/km/lane)		Speed (km/h)		Yield (km/h/veh)	
	Highway	Trunk road	Highway	Trunk road	Exact	Relative
Normal flow	15	33	110	60	76	100%
Highway impeded	20	33	90	60	71	96%
Trunk road impeded	15	40	110	45	63	81%
Both impeded	20	40	90	45	60	80%

This table is constructed like table C.1 on the preceding page, the difference is the flow is exchanged with lane density. The yield is the average travel distance for a vehicle per hour.
Contrary to table C.1 on the facing page there are no immediate undesired indications. Thus changing the measures has a beneficial impact on the usability of the aggregation.
The density is calculated based on a 2 second safety distance at the listed speed. This however does not necessarily reflect reality.

Table C.2: Independent measures

ods we turn our attention to the objective; the final construct based on our measurement framework that will be used to evaluate different solutions.

C.4 Objective

An objective is the goal to achieve. Unless otherwise mentioned we consider maximization – i.e. an *objective value* z of 10 is better than 9.

The objective can be considered in several ways. Here two alternatives are relevant. We can consider the system as a anonymous mass of vehicles and simply consider an aggregate over all vehicles. Alternatively we can consider each individual and then optimize on the distribution of the individuals.

In the first case – the anonymous mass – some possible objectives are:

- minimize the total trip time
- minimize the total travel distance
- maximize the average end-to-end velocity
- minimize the total congestion delay
- maximize trunk road flow

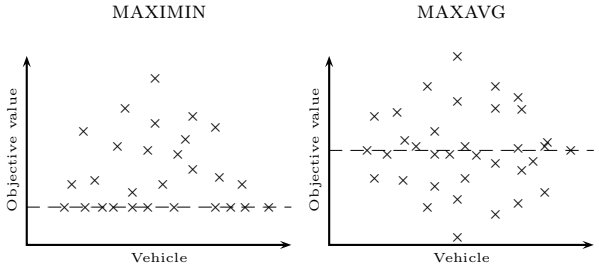
We call it anonymous mass representation because we have no interest in the individual vehicle. Mass representation solely focuses on a general aggregate. The equilibrium found by using a mass representative model is called the Nash equilibrium for a given criterion.

If we on the other hand wish to consider the individuals, there are several approaches from Mathematical Programming:

- MAXIMIN The idea is to maximize the minimum. The goal is to achieve the highest possible objective value for the individual with the lowest objective value.
$$\max \left\{ \min_{\forall veh} z \right\}$$
In less theoretical terms and in ABIT a MAXIMIN solution will ensure that no vehicle is individually given a particularily bad solution. The objective is constructed to optimize the worst solution for all individuals.
- MINIMAX The idea here is almost opposite the above. Theoretically it is stated as:
$$\min \left\{ \max_{\forall veh} z \right\}$$
In ABIT this has no relevance as this objective construct will ensure that all vehicles will be assigned to the worst possible paths.
- MAXAVG Here the idea is to get the highest average objective value.
$$\max \left\{ \text{avg}_{\forall veh} z \right\}$$
This solution is generally beneficial, but there is no attempt to ensure equal service to all vehicles. Some will be sacrificed to achieve a higher average.
- MINIMIN / MAXIMAX These alternatives are merely mentioned to complete the picture. It is clear that neither maximizing the maximum or minimizing the minimum makes any sense in ABIT. Both of these will lead to extreme differentiation between the vehicles. The MAXIMAX will send some vehicles on the highway and the rest on significantly worse routes. The MINIMIN will behave like the minimax mentioned above.

Figure C.2 on the next page shows the effect on the individual solutions when using MAXAVG or MAXIMIN. MAXAVG behaves almost like the mass aggregates and the major difference for the individuals is when using the MAXIMIN. We thus choose to consider mass aggregates – global objectives – and the MAXIMIN[3] – social objective – in this research project.

[3]and it's counterpart MINIMAX, when "less is better"



The figures show possible distributions of the individual objective values given the two constructs.

Figure C.2: Objective constructs

C.4.1 Global objectives

To add substance and clarity to the objectives the constraints on a feasible solution must be defined. These are:

- All vehicles must be moved from origin to destination
- Traffic regulation must be fulfilled

Neglecting either of these will lead to infeasible solutions.

To actually construct a model we have to specify the above constraints in greater detail.

First we introduce a *path* $p(A, B)$ from A to B, which is a set of sections, pieces of road, leading from A to B. The path must obey traffic regulations, such as one way roads, reserved lanes and turning restrictions. The path of vehicle i is thus $p_i(origin_i, destination_i)$. As the origin and destination is implicitly known for any specific vehicle this will be occasionally be shorthanded to simply p_i.

Given a pricing function $f(\cdot)$ we can denote the $f(\cdot)$-best path from A to B as $\widehat{p}^f(A, B)$. If $p_i = \widehat{p}^f(origin_i, destination_i)$ then vehicle i is said to have followed the $f(\cdot)$-best path.

C.4.1.1 Static Objectives

As we now have an outline formalism for a model we can define a group of simple global objectives. We call them static objectives as they are not influenced by the amount of traffic in the network. They can thus be solved efficiently with known All-pairs Shortest-paths algorithms or approximated efficiently with heuristics.

An example of a static objective is based on the distance measure $dist(p)$ of a path. The objective $\min\{\sum_{\forall i} dist(p_i)\}$ thus yields solutions where the total distance traveled is minimized. This solution can be found by ensuring that every vehicle follows its $dist()$-best path.

Unfortunately that solution is only attractive if all sections are traversed with similar velocities. Thus it may only be applicable to confined subregions in the network where the velocity can be assumed homogeneous.

In general the static objectives suffer from their inability to incorporate the flow on the individual sections of a path. They will propose the same route independent of the amount and distribution of vehicles in the infrastructure.

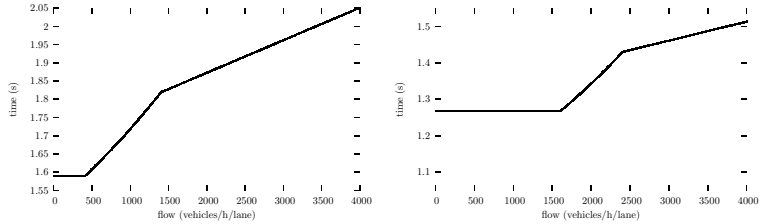
The impact of using a static objective may thus be catastrophic on all other objectives. This brings the dynamic objectives into consideration.

C.4.1.2 Dynamic Objectives

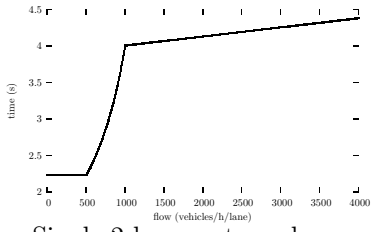
A dynamic objective is an objective that incorporates both amount and distribution of the traffic into the solution process.

A simple example thats depicts great complexities is the time measure $t(p)$. The objective $\min\{\sum_{\forall i} t(p_i)\}$ yields solutions where the total time in transit is minimized.

The problem is that the time measure of a section is dependent of the load of that section and it is even more complicated when considering intersections. An example from [Ortúzar and Willumsen, 2001] is the velocity/flow model of the Department of Transport in the UK from 1985. The model has four parameters; two velocity measures S_1 and S_2 and two flow measures F_1 and F_2. The variable



Single 2 lane, rural Dual 2 lane, rural



Single 2 lane, outer urban

The figures depict three examples of roads according to equation C.1. Notice that the two graphs on the top have a different time scale than the one below. As expected the time to traverse a section is static while the traffic is low. This is called the free flow time. Once the free flow capacity is exceeded the traversal cost increases rapidly until the congestion capacity is reached. Beyond congestion capacity the traversal time still increase, but much slower than during the onset of congestion – between free flow capacity and congestion capacity.

Figure C.3: Time/flow relationship for different road sections

F is the flow into the section of length d.

$$T(F) = \begin{cases} \frac{d}{S_0} & F < F_1 \\ \frac{d}{S_0 + (F_1 - F)\frac{S_0 - S_1}{F_1 - F_2}} & F_1 \leq F \leq F_2 \\ \frac{d}{S_1} + \frac{F}{8(F_2 - 1)} & F > F_2 \end{cases} \tag{C.1}$$

Figure C.3 shows some plots of the model. It is clear that the time measure exhibits heterogeneity over different, but familiar section types. Only when the flow is relatively low is it possible to apply static solution methods.

Other models for flow/time relationship, like the 1964 study by the Bureau of Public Roads (USA) in [Ortúzar and Willumsen, 2001], are smooth, but not homogeneous. The parameters required are the practical capacity of the section

Q_p, free flow time T_0, and the calibration parameters α and β.

$$T(F) = T_0 \left(1 + \alpha(F/Q_p)^\beta\right)$$

It is possible that the solution methods will be more efficient with smooth than piecewise linear relationships.

C.4.2 Social objectives

As with global objectives we can divide these into either static or dynamic. An important observation here is that for the static objectives the global and social version will both yield the same solution. This is because there is no social sensitivity in the static objectives as the distribution of other vehicles do not influence on the optimal route for any single vehicle.

We thus have to focus on the dynamic social objectives. This time we choose the fuel consumption measure $fc()$. This is dynamic as consumption is dependent both on the distance, the time and the fluctuations of the velocity through the network.

The social fuel consumption objective is:

$$\min \left\{ \max_{\forall i} fc(p_i) \right\}$$

Applying this directly is not necessarily a good idea. The problem is that we only consider the total individual measure. Driving 10 kilometers obviously requires more fuel than driving only 1 kilometer. Any solution method will thus sacrifice all short routes to ensure optimal consumption for the longer routes.

An approach to this problem is to use the relative social objective:

$$\min \left\{ \max_{\forall i} fc(p_i)/fc_(p_i^{free}) \right\} \tag{C.2}$$

or the absolute social objective:

$$\min \left\{ \max_{\forall i} fc(p_i) - fc_(p_i^{free}) \right\} \tag{C.3}$$

where p_i^{free} is the free flow – alone-on-the-roads – path for vehicle i. The difference in the two is that the relative social objective penalizes the expensive routes the most. The absolute social measure penalizes all with the same amount. Thus, if the long route uses and extra liter of gas so could the short

route. Driving a detour of 10 kilometers for a 3 kilometer trip is significantly more annoying than extending your 100 kilometer trip by another 10 kilometers.

I believe that it will be more acceptable to the public if the relative social objective is used.

C.4.3 Selected objectives

All of the above constructs will be considered. Dynamic objectives are evidently necessary to produce usable results. The relative social objectives will possibly produce significantly different and more acceptable solutions. Finally, the static objectives as all objectives are static when considering free flow for a single vehicle.

The measure that we will focus on will be time. The reason for this is that it, as shown in the examples above, exhibits very complicated dynamics and that it is well covered in scientific literature.

C.5 Conclusion

Measuring traffic contains many aspects. Here we have chosen to consider time as main measure. This was chosen because Discrete Choice Modeling has shown that time is essential to the drivers, time/flow relationships is being researched intensively these days and finally because the relationship functions between time and flow present a significantly new approach in Operations Research (OR). Three objective constructs were discussed: the static objectives, the dynamic objectives and the relative social objectives. The three objective functions chosen in these categories are:

- The static time objective

$$\min \left\{ \sum_{\forall i} t(p_i^{free}) \right\}$$

- The dynamic time objective

$$\min \left\{ \sum_{\forall i} t(p_i) \right\}$$

- The relative social time objective

$$\min \left\{ \max_{\forall i} t(p_i)/t(p_i^{free}) \right\}$$

All subject to the physical and political laws of traffic.

All these objective function constructs should be properly compared on a suited test set.

Bibliography

[Abdulhai and Look, 2003] Abdulhai, B. and Look, H. (2003). Impact of dynamic and safety-conscious route guidance on accident risk. *Journal of Transportation Engineering*, 129(4):369–376.

[Bovy and Stern, 1990] Bovy, P. H. L. and Stern, E. (1990). *Route Choice: Wayfinding in Transport Networks*. Kluwer Academic Publishers.

[Ortúzar and Willumsen, 2001] Ortúzar, J. D. D. and Willumsen, L. G. (2001). *Modelling Transport*. John Wiley & Sons Inc., 3rd edition.

[Roughgarden and Tardos, 2002] Roughgarden, T. and Tardos, E. (2002). Operations research - how bad is selfish routing? *Journal of the ACM - Association for Computing Machinery*, 49(2):236–259.

Appendix D

Quasi Newton Method used for TAP with inseparable cost function

Abstract: We present a new inseparable cost function and an augmented solution approach. The cost function can be calibrated to accomodate the extra delay in turning movements caused by right of way traffic and the solution method guarantees convergence in quadratic time even for inseparable cost functions.

Although Bar-Gera's Origin-Based Algorithm (OBA) for the Traffic Assignment Problem (TAP) apply almost the same algorithm, we differ in significant areas. We do not use approach proportions and our projected Quasi-Newton Method (QNM) is not restricted to a diagonal Hessian. Furthermore our solver is based on two different software packages and not tightly integrated.

At a significant performance impact this augmented solver can solve the same instances. The usage of the full Hessian only requires the cost function to be locally polynomial and not separable.

We combine a true QNM with generic network algorithms to achieve a much more versatile solver for the TAP. This solver is then used to illustrate scenarios related to disruptions and the application of an inseparable cost function.

Keywords: Traffic Assignment Problem, Quasi-Newton Method, route generation, inseparable cost function

D.1 Introduction and background

In traffic science the Traffic Assignment Problem (TAP) is crucial in many applications. In most cases only the flow on different sections are interesting which has led to solvers that do not retain the origin-destination route information between iterations or in the solution.

In our case we solve the TAP with a distinct requirement. The project Agent Based Individual Traffic guidance (ABIT) described in [Wanscher, 2004] and [Wanscher, 2006] requires the individual routes of all vehicles to be known and this automatically enforces us to retain the route information.

It has been shown by [Bar-Gera, 2002] that retaining and using this information significantly speeds up the convergence time for the solution. [Boyce et al., 2004] states that actually finding the optimum solution is unnecessary. The key issue is to identify the necessary convergence criterion. Identifying the inter iteration solution value gap for a stable solution is important. Once this is found the solution process can be stopped and a suitable solution used. Thus the convergence rate to this gap is more important than absolute convergence.

Applying the Origin-Based Algorithm (OBA) directly however is not sufficient. It only guarantees convergence for separable cost functions. As we shall see later a much more micro realistic cost function is necessary for ABIT.

D.2 Agent Based Individual Traffic guidance

Agent Based Individual Traffic guidance (ABIT) is the next evolutionary level of GPS-enabled in-vehicle routeplanning devices. Today the Traffic Message Channel (TMC) [Forum, 2004] in Europe is broadcasting incidents in the road infrastructure to TMC enabled GPS-units, which then can act accordingly.

The next step is to make the individual GPS-units actually communicate back to a system. This could then be used to automate the broadcasts over TMC instead of the manual interaction mostly required today. The Honda car manufacturer has included such a two-way system in their vehicles sold for Japan.

The drawback, however, is that is it still left to the individual GPS-unit to find a good path.

The main concept in ABIT is to centralize this and then redistribute the flow of traffic based on the destination of every single vehicle as well as immediate and long term throughput conditions.

The amount of data is devastating and organisation and distribution of it is an onerous task. However in this article we focus on the actual redistribution of flows given realtime information and models. In ABIT terminology defined in [Wanscher, 2004][1] we are in this paper considering a solver for the immediate dispersion.

D.3 A More Precise Cost Function

Given the online nature of ABIT it is clear that the usual assumption on the TAP section cost function given in equation D.1 is to coarse.

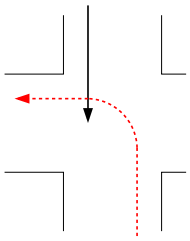
$$T_{\overrightarrow{od}}(|\overrightarrow{od}|) = fft_{\overrightarrow{od}} \left(1 + \alpha \frac{|\overrightarrow{od}|}{\left\lceil \overrightarrow{od} \right\rceil \cdot \gamma_{\overrightarrow{od}}} \right)^{\beta} \tag{D.1}$$

In this equation fft is the freeflow time, γ is the green period fraction for light controlled intersections and α are β are per section calibration parameters. β is in numerous cases set to 4. It assumes that the traversal cost is separable over every single section – the cost of traversing a section is only dependent on the flow on that section. In many cases this is sufficient as the aim is to find an estimated aggregate on flow distribution for strategic purposes. The immediate detail of any single intersection is thus irrelevant as the network is usually viewed over a longer period of time.

In ABIT however it will be necessary to have the much more detail of the immediate network behavior. On the other hand it is not viable to keep extending the model as the solution time is severily increased for every included complexity. However, discarding models with long solution times, but high quality solution is unwise. Although the model and the solution method itself may be unusable they may lead to fast methods of acceptable quality.

In research concerning congestion the focus is on mainly spillback. This happens when the congestion on a single section generates queueing beyond the section

[1] included here in appendix A



The flow depicted by the black path has right of way. The cost of following the red path is thus dependent on the load on the black path.

Figure D.1: Example of right-of-way situation

which is congested. The queuing then extends back on the sections which lead to the congested section – hence spillback. This effect is primarily essential when the network is heavily loaded.

In our case we need greater detail for networks of any load as the system is ment to be constantly online. Although it will be possible to exchange the underlying solver for the different cases we will at this point aim at a versatile solver.

The issue we focus on is the extra cost associated with crossing through traffic. Figure D.1 shows an example where the usual aggregation is insufficient. The flow on the turning movement is impeded by the opposing flow going straight through the intersection. Any separable cost function is unable to reflect this dependency. This specific dependency – that the cost of traversing a section is also influenced by the flow on another section – is inseparable.

We propose an inseparable cost function, which can be calibrated to accomodate interfering flows.

$$T_n'() = \left(T_n\left(f_n\right) + U_n\left(f_n, \sum_m y_{nm} \cdot f_m\right)\right) f_n \tag{D.2}$$

Here n and m represents sections such as $\overrightarrow{ab}$ or $\overrightarrow{de}$.

The formulation is an extension to equation D.1, which includes the inseparable constituents as linear dependencies.

The U_n function is a gap acceptancy function (GAF) and we adopt the following simplistic formulation for this paper:

$$U(f_n, f_m) = \left(e^{f_n f_m} - f_n f_m - 1\right) f_n^{-1} \tag{D.3}$$

In this case the impact of yielding is represented by the product of the right-og-way flow and the yielding flow. With no right-of-way flow there is no impact.

Obviously increasing the right-of-way flow has higher impact than increasing the yielding flow. The functions here are not fully calibrated, but are loosely estimated to illustrate the impact on the solver and the solution.

It is possible to include spillback reflection in the cost function, but the intent here is to shed light on inseparability, not construction of the most complicated cost function.

D.3.1 The model

In ABIT we consider the social or Nash equilibrium. However in this case we compare our solver with the OBA solver [Bar-Gera, 2002] ising the traditional formulation based on the Beckmann transformation [Beckmann et al., 1956].

We define the integer variable p_{ijk} as the number of vehicles on the k^{th}-path from node i to node j. In total there are K_{ij} routes from i to j. D is the demand matrix where D_{ij} is the amount of vehicles from i destined at j. a^n_{ijk} is 1 if p_{ijk} uses section n and 0 otherwise. c_n is the estimated capacity of section n and finally T_n or T'_n is the cost function given in equations D.1 or D.2. f is the aggregated flow vector, which at each n^{th} element has the total flow on section n: $f_n = \sum_k^{K_{ij}} a^n_{ijk} p_{ijk}$.

The model is then:

Minimize	$\sum_{n,i,j} (T'_n(f))$	Social equilibrium
Subject to:	$\sum_k p_{ijk} \geq D_{ij}, \forall i, j$	Demand satisfaction constraints.

This is essentially a network flow problem where the capacity constraints are omitted. The capacity of a section is in this case included in the model through the cost function, which includes the theoretical capacity or other parameters to estimate the cost of assigning a given flow to a section. If the problem is a minimization problem and the cost functions are monotonically increasing there is no need to introduce explicit capacity constraints.

D.3.2 Extensive Representation

Representing the inseparable real world effects can in sometimes be directly included in the more complicated cost function. However some situations are more properly addressed differently. The example we choose here is with yielding traffic in an ordinary intersection.

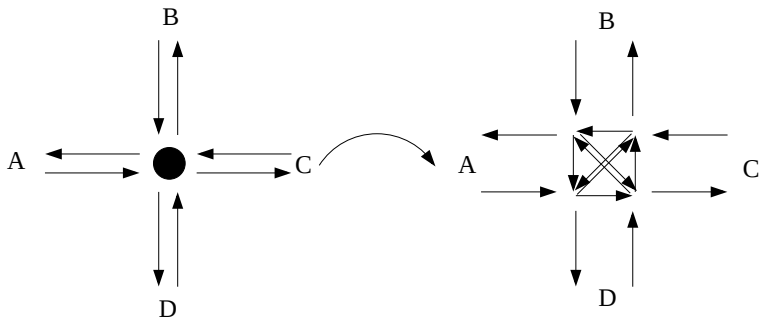


Figure D.2: Small example network

Consider the small example network depicted in figure D.2. Representing the GAF by the y_{nm} parameters in equation D.2 is possible, but to coarse. The simplication that all traffic from B is dependent on traffic from D might not be desirable. Evidently turning traffic from B to C is directly depent, but traffic from B to A or D is only indirectly affected by traffic from D. This is when queuing turning traffic from B to C prevents traffic from B to A or D from passing. The y_{nm} variable for straight through and turning traffic could thus be different and this differentiation could be included by additional nodes and sections as depicted on the figure.

In our solver, however, we can directly integrate this into the path generation and we do not need to increase the number of sections or nodes. The path generation and cost calculation time is increased as the cost of a given path is more complicated to calculate. Most importantly neither the number of nodes nor the number of sections are increased.

D.4 The solver

The TAP problem has been researched for several decades. Originally it was proposed as above, but it was quickly realised that the number of possible paths in a network grows exponentially with number of sections and intersections.

Frank and Wolfe first proposed an actually usable solution method now known as the Frank-Wolfe method [Frank and Wolfe, 1956]. The idea is to continously redistribute smaller fractions of the traffic until the fraction is sufficiently small or the solution value satisfies a given convergence criterion.

It required many iterations and, for that time, large amounts of memory which led to the new flow based formulation of the problem. The solution to this problem was significantly easier to find, but still suffered from memory consumption and slow convergence.

Within the last decades several publications – [Jayakrishnan et al., 1994], [Chen et al., 2002],[Bar-Gera, 2002] and [Bierlaire and Crittin, 2006] – have proven significant speedups by returning to the path based approach, but using a different redistribution of the flow.

The most promising of these is the OBA in [Bar-Gera, 2002] previously mentioned. It uses an approximation of the Quasi-Newton Method (QNM) described in [Davidon, 1959], [Fletcher and Powell, 1963] and [Broyden, 1965] in the field of mathematics. Bar-Gera's approximation to a diagonal Hessian is valid as he only consideres separable cost functions. Our application of inseparable cost functions however will not be properly represented in the diagonal of the Hessian. Only the separable parts of a function is represented in the diagonal. The full potential of the quasi-Newton Method is not utilized until the full Hessian is approximated by the gradient vector.

Our solver is constructed from the Goblin [Fremuth-Paeger, 2007] network algorithm package and the Minos [Saunders, 2006] Fortran solver.

The reason for not specifically optimizing the solvers into one is that we wish to show that applying state-of-the-art algorithms from mathematics and operations research is sufficient for making an efficient solver. The interaction between the solvers has of course been optimized.

D.5 Quasi-Newton Projection

In the conventional Frank-Wolfe method for TAP the idea is to continuously redistribute a smaller fraction of the flow until convergence is reached. This is done by iterations that first find the set of all points shortests paths and then redistribute a fraction of the flow to these shortest paths. By reducing the fraction for each iteration the solution converges to the intended equilibrium.

The first adaptations of this method however converged slowly. This was due to residual flows on non-optimal paths, which only very slowly would be moved to optimal paths. Furthermore they required large amounts of memory and at the time only small instances could be addressed.

The later articles all try to circumvent this by using smarter and more intelligent redistribution of the flows and path generation. In OBA every generated path from an origin to a destination is stored. The redistribution of flow is then done by a simplified QNM also called a Pseudo-Newton Method (PNM).

The essential part of the Pseudo-Newton Method is that the curvature of the solution space is approximated by a numerically generated reduced diagonal Hessian. The generation of the approximated Hessian is done by iterative updates. In every iteration the Hessian is reestimated based on current and previous iterations. After several iterations the approximated Hessian is sufficiently precise to substitute the exact Hessian for this application. This method is theoretically proven to be very efficient for this specific type of problem.

In our case we do not restrict the Hessian representation to the diagonal. We wish to show that using the full Hessian maintains convergence even for inseparable cost functions as long as the objective function is locally polynomial and the solution space only contains one minimum.

Our solver is constructed much like the traditional approaches to the TAP.

1. Set all flows on all sections to zero and the size of the paths pool to zero.
2. Find all pairs shortests paths.
3. Add all paths found not already in the pool to the pool.
4. If no new paths were added to the pool go to step 8.
5. Distribute the demand flow with the QNM with the path pool.
6. Set the flows on all sections according to the found distribution.
7. Go to step 2.
8. The path pool and the distribution found by the QNM solver is the best solution for the predefined convergence criterion.

In the above step 5 is handled by MINOS and step 2 is done through GOBLIN.

During the writing of this article [Bierlaire and Crittin, 2006] presented a very interesting article describing an even more advanced approach to solving large-scaled noisy fixed-point problems. They present a generalization of secant methods, and uses several iterates to generate linear approximations. The method belongs to the Quasi-Newton family of methods, but their approach is matrix free, thus allowing them to solve large-scale systems of equations. Furthermore

the method is not dependent on the existence of derivatives of the equations and there is no particular assumption on the problem structure or the problems Jacobian. However their method still has some limitations, but the direction is promising.

D.6 Experimental results

We chose two specific instanses the SiouxFalls and the ChicagoSketch network from the standard TAP library. For each of these we modified the demands or the section capacities to test the solver. Some demands are altered specifically to introduce diversity in path generation and as such requires far more solution time.

The SiouxFalls instance consists of 24 nodes and 76 sections. There are 528 origin-destination pairs with non-zeros demand.

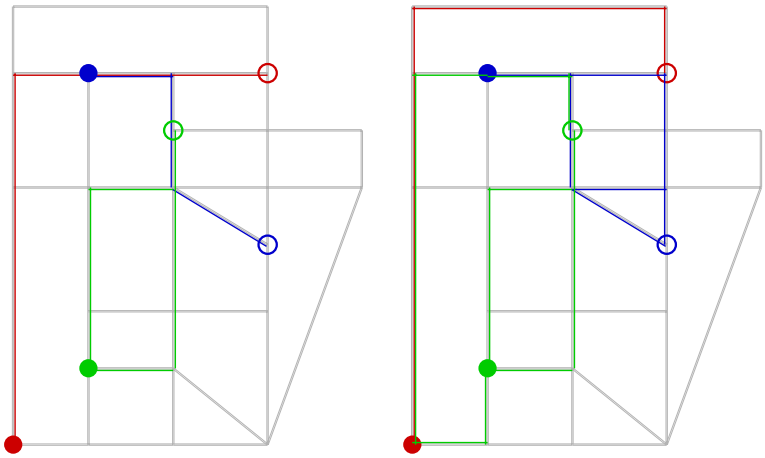
The ChicagoSketch instance consists of 546 real nodes, 387 centroids (virtual demand nodes) and 2950 sections. For this instance the number of non-zero demand pairs is 93513.

All tests were performed on the same machine running either Windows XP for the OBA solver or Kubuntu linux for our solver.

D.6.1 Gap acceptancy function

The original intent of this paper was to use the GAF given in equation D.3. However this function is not easily integrated. We applied a simplistic numerical approximation of the integral. Even if we increased the running time of the solver four fold by refining the integral value the noise generated by the approximation confused the QNM solver. The performance was significantly deteriorated and the path generation also suffered from the erratic behavior of the solver. The table below shows the running times using equation D.3 compared to a simpler inseparable multiplicative approximation to the GAF.

Instance	Time	Noise warnings
SiouxFalls with complex GAF	$0.38s$	26
SiouxFalls with linear GAF	$0.2s$	0



Traditional cost function Yield cost included

The above figure compares the same two instances of the SiouxFalls network. On the left the traditional separable cost function is used and on the right our inseparable cost function is used. Three of the o/d pairs and their paths have been highlighted. The filled and empty circles are respectively the origins and destinations of the paths of the same color. Both the green and the blue pair move some flow to paths with only right turns. The path change for the red pair is done to minimize right-of-way traffic for other flows.

Figure D.3: The impact of yielding

We thus chose the simple GAF for our experiments on the inseparable cost function. Figure D.3 shows the change in a solution, when a GAF is used in the model.

D.6.2 Runtime analysis

Instance	SiouxFalls			ChicagoSketch	
Solver	QNM		OBA	QNM	OBA
Cost function	separable	inseparable	separable	separable	separable
Average total time	1.23s	0.22s	0.64s	39717s	148s
Average flow time	1.16s	0.20s	0.62s	37091s	138s
Average flow fraction	94.3%	90.9%	96.8%	93.4%	93.2%
Average path time	0.04s	0.02	0.02s	2544s	10s
Average path fraction	3.2%	9.1%	3.1%	6.4%	6.8%

In the table above the times for all SiouxFalls and ChicagoSketch instances are shown. The yield matrix for the inseparable cost function for the SiouxFalls

instances has been constructed manually. Due to the size of the ChicagoSketch instance no corresponding yield matrix has been constructed yet.

It was unexpected that the performance for the instances with the GAF outperformed the others. Apparently the simplistic GAF we apply improves the reduction in the residual flow on inefficient paths.

Given the average fraction of time used for the flow distribution it is clear that mathematical improvements will be reflected in the total time. On the other hand the time spent on startup, path generation and administration is so small that adding even complicated, but well scaling improvements, might reduce the work for the flow distribution considerably.

It is however also clear that the combination of chosen path generator and QNM solver scales unacceptably much worse than the OBA approach used in [Bar-Gera, 2002]. It is unlikely that this combination will be usable for large scale networks, but for small networks the performance of the solver is sufficient for a real-time application.

D.6.3 Convergence

The graphs in figure D.4 illustrates that approximately 70% of the solution time is used to achieve convergence at a very slow rate. However, as we are considering autonomous vehicular traffic, absolute convergence might not even be necessary. And as the solutions are very close we may reduce the running time limit significantly and still get applicable results.

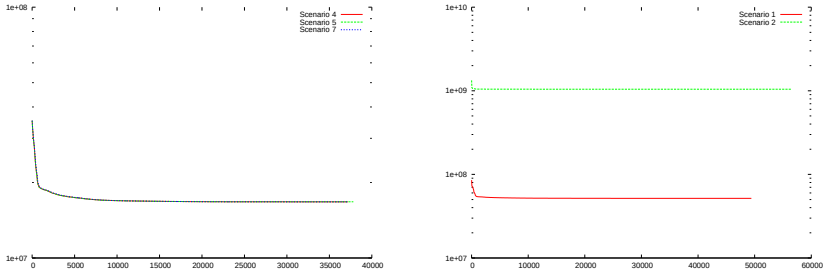
D.7 Conclusion

Introducing a more complicated cost function is expected to increase the convergence time. However the recent research presented apparently alleviates this to some extent by applying mathematical findings properly to the TAP, Dynamic Traffic Assignment (DTA) and the Consistent Anticipatory Route Guidance (CARG). The experimental results presented in [Bierlaire and Crittin, 2006] shows that their approach is interesting not only due to versatility, but also based on efficiency and convergence rate.

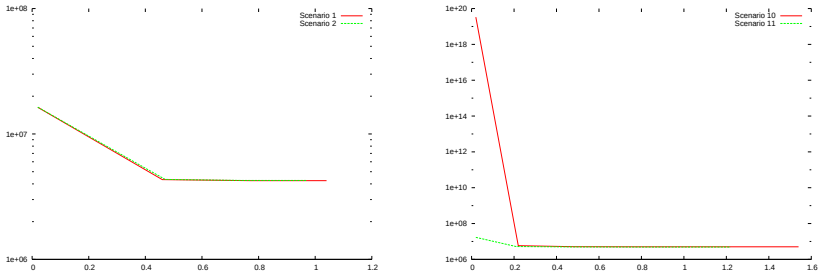
Our findings indicate that optimized path generation might lead to more efficient solvers. Further more the effect observed by inserting traffic and introducing

Instances with light disruption Instances with significant disruption

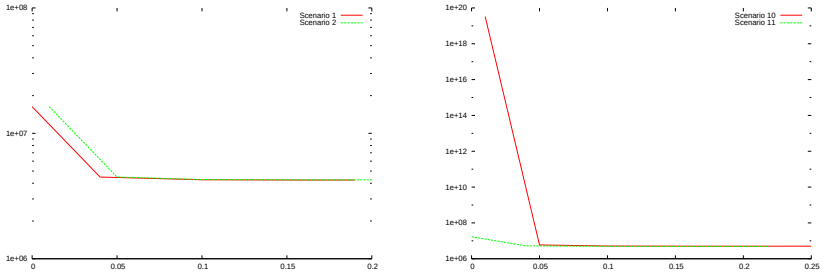
ChicagoSketch



SiouxFalls



SiouxFalls with gap acceptancy function



Each of the graphs above depict the relation between alloted solution time and best found solution. The initial iterations rapidly reduce the objective value, but as expected the objective value stabilizes and multiple iterations are necessary to acheive just small advances.

Figure D.4: Convergence for solver

disruptions in the network was illustrated.

Even though these mathematical methods remove some of the necessary requirements for guaranteed convergence they still do not alleviate the impact of network expansion. However it might be possible only to increase the number of sections and nodes and not the number of generated paths directly. The time consuming QNM solver is thus not encumbered by such an expansion.

As ABIT is concerning a stochastic autonomous mass of individuals it might be possible through statistics to reduce network expansion thus reducing the number of variables. This could be done by identifying the insignificant introductions of inseparability in the cost function and then remove the variables that represent this.

In the case of ABIT and concerning a stochastic autonomous mass of individual drivers it is most likely not necessary to actually find the optimum. If the convergence criteria is slacked the execution time can be reduced significantly dependent on the desired precision.

Bibliography

[Bar-Gera, 2002] Bar-Gera, H. (2002). Origin-based algorithm for the traffic assignment problem. *Transportation Science*, 36(4):398–417.

[Beckmann et al., 1956] Beckmann, M., McGuire, C., and Winsten, C. (1956). *Studies in the Economics of Transportation*. Yale University Press, New Haven, Conneticut.

[Bierlaire and Crittin, 2006] Bierlaire, M. and Crittin, F. (2006). Solving noisy, large-scale fixed-point problems and systems of nonlinear equations. *Transportation Science*, 40(1):44–63.

[Boyce et al., 2004] Boyce, D., Ralevic-Dekic, B., and Bar-Gera, H. (2004). Convergence of traffic assignments: How much is enough? *Journal of Transportation Engineering*, 130(1):49–55.

[Broyden, 1965] Broyden, C. G. (1965). A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation*, 19(92):577–593.

[Chen et al., 2002] Chen, A., Jayakrishnan, R., and Tsai, W. K. (2002). Faster frank-wolfe traffic assignment with new flow update scheme. *Journal of Transportation Engineering*, 128(1):31–39.

[Davidon, 1959] Davidon, W. (1959). Variable metric method for minimization. A.E.C. Research and Development Report ANL-5990 (Rev. TID-4500, 14th ed.), A.E.C. Research and Development.

[Fletcher and Powell, 1963] Fletcher, R. and Powell, M. J. D. (1963). A rapidly convergent descent method for minimization. *Computer Journal*, 6:163–168. MR 23, 2096 (1964).

[Forum, 2004] Forum, T. (2004). Tmcforum.com. Online by TMC Forum. direct link http://www.tmcforum.com/.

[Frank and Wolfe, 1956] Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110.

[Fremuth-Paeger, 2007] Fremuth-Paeger, C. (2007). Goblin: A graph object library for network programming problems. Internet site. http://www.math.uni-augsburg.de/ fremuth/goblin.html.

[Jayakrishnan et al., 1994] Jayakrishnan, R., Tsai, W. K., Prashker, J. N., and Rajadhyaksha, S. (1994). Faster path-based algorithm for traffic assignment. *Transportation Research Record*, 1443:75–83.

[Saunders, 2006] Saunders, M. (2006). Minos 5.51. Internet site. http://www.sbsi-sol-optimize.com/asp/sol_product_minos.htm.

[Wanscher, 2004] Wanscher, J. B. (2004). Agent based individual trafficguidance. In *Trafikdage 2004 (http://www.trafikdage.dk/)*.

[Wanscher, 2006] Wanscher, J. B. (2006). Agent based individual traffic guidance. In Hansen, L. G., Nielsen, L. D., Nielsen, O. A., and Hels, T., editors, *Artikelsamling 2006*. Trafikforskningsgruppen, Aalborg Universitet.

Appendix E

Further Results

The solver developed in Quasi-Newton Method for TAP with inseparable cost function in the previous chapter was also used to generate the following results. They support the expected bending of the flows when disruptions occurs. Furthermore they exemplify the impact on the paths considered and used by the solver for different Level of Influence (LoI).

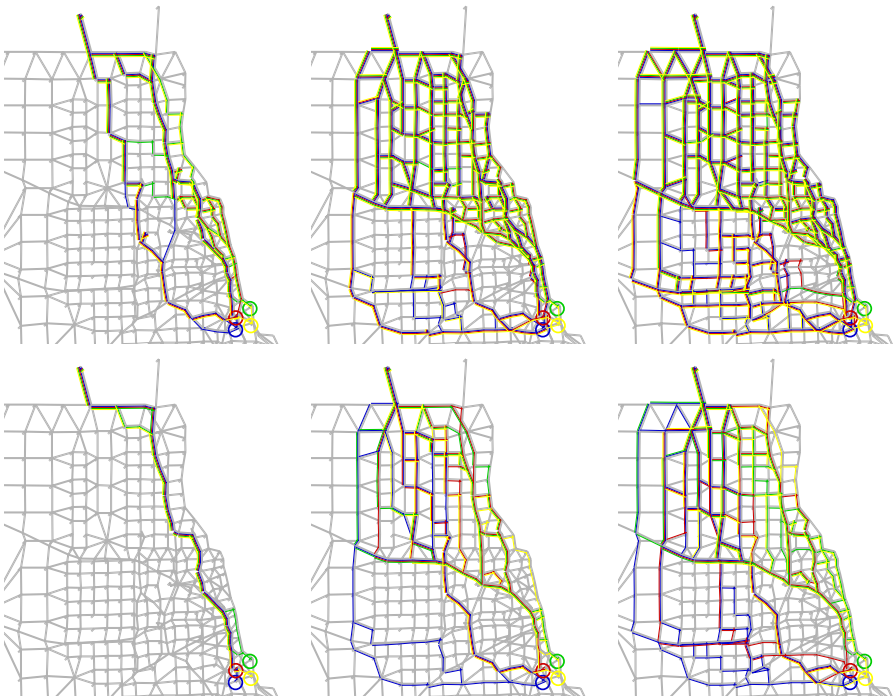
E.1 Path generation

When examining the path generation both number used and unused of paths are interesting. Further more inspecting the diversity yielded an interesting result. The diversity is the number of generated or used paths per origin-destination pair. From table E.1 above it can be seen that for the SiouxFalls instances approximately twice the amount of necessary paths are generated. In the ChicagoSketch instance this deficiency is even worse with generating three to six paths for every single one used. This indicates that better path generation might improve performance as the time spent in the solver might be reduced.

By increasing the pressure in the network, either by introducing excess demand or by reducing the capacity of specific sections, we made two observations from the numbers in table E.1:

Per o/d pair	SiouxFalls	with extra demand	ChicagoSketch	with extra demand
Avg generated paths	2	2.18	3.04	6.54
Avg used paths	1.06	1.06	1	1
Usage	52.8%	48.7%	33.0%	15.3%
Max generated	5	6	32	360
Max used	3	3	3	16
Time used	0.9s	1.33s	38147s	56473s

Table E.1: Paths generated and used for the two instances

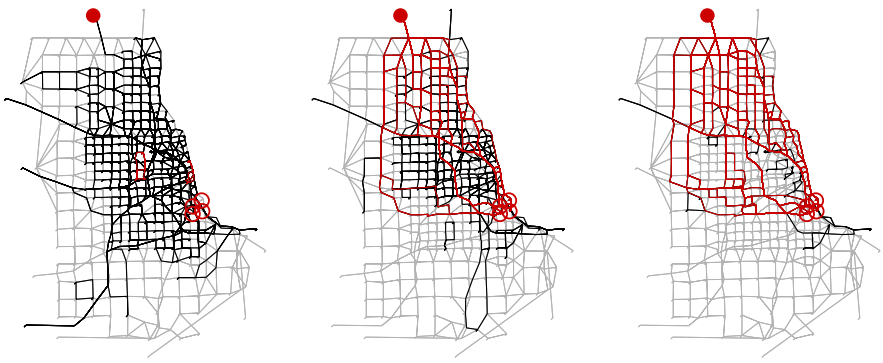


The illustration in the top row show the increase in generated paths for three modified ChicagoSketch instances. The bottom row shows the used diversity for the same instances. The demand is least in the left column and most in the right column. The paths are colorcoded according to their distination point.

Figure E.1: Generated and used diversity

- The number of generated paths increased
- The number of used paths per o/d-pair decreased

The increased in generated diversity was expected as it would be more difficult to find all suitable paths. However the decrease in used diversity was unexpected.



a) b) c)

a) shows the ChicagoSketch with no demand alteration. On b) the demand has been intensified and on c) the network is practically flushed from the designated origin. The black sections are sections which are used by o/d-pairs that splits the flow between two paths. The red sections used by o/d-pairs that uses three or more different paths. The filled red circle is the origin where the flow is inserted and the outlined circles are the destinations.

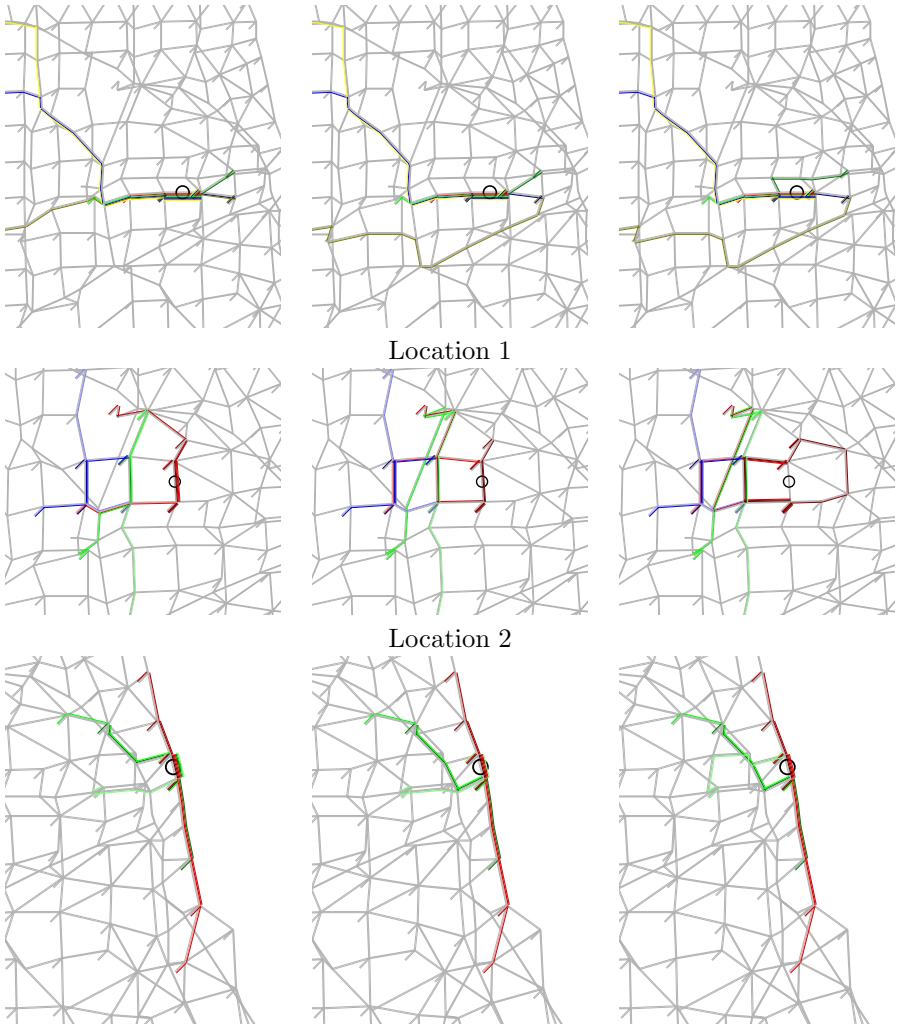
Figure E.2: Diversity

The solutions shown in figure E.2 shows that for the directly affected o/d-pairs the usage diversity is increased. However for almost all other pairs the usage diversity is decreased as all the vehicles from same o/d-pair use the same, but to some extent, different path.

For the ChicagoSketch instances the generated diversity was expectedly higher for the intensified instance. 608782 paths were generated for this opposed to only 283104 paths for the lightly loaded instance. The intensified instance used only 182 paths less than the 93390 used for the light instance. In the intensified instance 93118 o/d-pairs used only a single path for the entire demand. This is 226 more than the standard instance. The concentration affects 0.2% of all pairs in our scenarios.

E.1.1 Flow bending

The scenarios depicted in figure E.3 clearly shows how the flows bend to accomodate the increased traffic pressure in relation to a disruption.



Location 1

Location 2

Location 3

The sequences above show the impact on of the paths used in the solution if thecapacity of a section is reduced. The left most column shows that undisrupted network. And the figures on the right show the effect of decreasing the capacity gradually.

Figure E.3: Bending flows

Appendix F

Drawing a random number

Jørgen Bundgaard Wanscher
Majken Vildrik Sørensen

Kgs. Lyngby 2006
IMM-TECHNICAL REPORT-2006-2

Abstract: Random numbers are used for a great variety of applications in almost any field of computer and economic sciences today. Examples ranges from stock market forecasting in economics, through stochastic traffic modelling in operations research to photon and ray tracing in graphics.

The construction of a model or a solution method requires certain characteristics of the random numbers used. This is usually a distribution classification, which the sequence of random numbers must fulfill; of these some are very hard to fulfill and others are next to impossible. Today mathematics allows us to transform distributions into others with most of the required characteristics. In essence, a uniform sequence which is transformed into a new sequence with the required distribution. The subject of this article is to consider the well known highly uniform Halton sequence and modifications to it. The intent is to generate highly uniform multidimensional draws, which are highly relevant for todays traffic models.

This paper shows among others combined shuffling and scrambling seems needless, that scrambling gives the lowest correlation and that there are detectable differences between random numbers, dependent on their generation.

Keywords: Mixed Logit estimation, *a priori* distribution, random numbers, Halton numbers, Shuffled Halton, Scrambled Halton, Leaped Halton.

F.1 Introduction and background

A mixed logit model is a generalised member of the logit family, where one/ some of the coefficients are no longer scalars; but rather stochastically distributed variables. Over recent years increased focus has been put on how these variables could be described (which *a priori* distribution should be applied, could this be estimated, could this be tested, etc).

Mixed logit models, ie. models including distributed components are today at the frontier of application and development in transport modelling. The name Mixed Logit has for some years been used indiscriminately with Error Component Models, Random Parameters Logit, Models with Stochastic (Distributed) Preferences (Coefficients), Logit Kernel Models or Hybrid Choice models, for models where error components are added to the traditional (linear) utility function. However, the name conventions Mixed logit models seems to have taken the lead.

During the past 5 years the use of such models is growing rapidly, due to an increased access to software (both commercially by Hague Consulting Group (Alogit4ec), the self-contained BioGeme [Bierlaire, 2002] and shareware as Gauss code [Train, 2002]. In general, the method of Maximum Simulated Likelihood (MSL) is applied, although this only optimises within a given *a priori* distribution of the error components.

When mixed models are estimated or used for forecasting, emphasis must be put on how the stochastically distributed terms are obtained as this may significantly impact the model results. The interesting question of correlation between these error components, or the *a priori* assumption of shape of distribution has been dealt with in [Sørensen, 2003a, Sørensen, 2003b, Sørensen, 2004]. Yet an unresolved questions is that of how to generate the distributed terms.

This paper will focus on *how* random numbers can be generated, and the differences between these. First, the theoretical framework is defined in section

F.2, different methods for single dimensional draws will be covered in section F.3, followed by multidimensional draws in section F.4. Following this runtime analysis in section F.6 will preceed comparison of the numbers generated by means of correlation in section F.7. This paper is concluded in section F.8.

F.2 Theoretical framework

In the following a general model is formulated, for which the empirical distribution of the stochastic coefficients in an error components model can be determined. Let the utility function be specified as

$$U_{ji} = V_{ji} + \nu_j \tag{F.1}$$

where V_{ji} respective ν_j represents the explained respective the unexplained variation in the utility function for choice of alternative i by individual j. The explained variation is often formulated as a linear relation, i.e. $\sum_k \beta_k X_{jik}$ where each attribute X_k is converted into an arbitrary measure *utility*. This framework is (usually) operationalised by assuming that the unexplained part of the variation is described by some well-known distribution (Extreme Value, type I (EVI) leads to the logit formulation). Estimation is performed by means of maximum likelihood, where the estimates of the coefficients are conditional on the specification of the utility function (which potentially transformed attributes are included) as well as the *a priori* (assumed) distribution of the ν.

The expansion to the mixed logit framework[1] corresponds to partitioning of the unexplained variation (the ν) into a *systematic* (describable) and an *unsystematic* part. Formulated as in (F.1)

$$U_{ji} = V_{ji}(X_{ij}) + \eta_j f(X_{ij}) + \epsilon_j, \tag{F.2}$$

where η_j is the systematic part of the unexplained variation, while ϵ_j is the unsystematic part. Both the deterministic part of the utility and the systematic part of the utility may be functions of the explanatory variable (X_{ij}), while the unsystematic part (ϵ) is (again) assumed to be EVI distributed; hence, estimation by means of the logit framework is facilitated.

[1]The 'Mixed Logit Framework' refers to the Mixed Logit Model ([McFadden and Train, 2000]), Error Components Logit ([Ben-Akiva et al., 1993]), Logit Kernel ([Ben-Akiva et al., 2001]), Random parameters Logit ([Ben-Akiva et al., 1993]) and the Hybrid Choice model [Ben-Akiva et al., 2002].

The elements of the matrix ξ are either assumed to follow a stochastic distribution with mean zero and variance σ_j^2 or are identical 0 (fixed coefficients). Traditionally distributions suggested are the normal ([McFadden and Train, 2000]) and the lognormal ([Ben-Akiva et al., 1993], [Train, 1998]), though other distributions have been suggested which include the χ^2 ([Nunes et al., 2001]), the uniform ([Revelt and Train, 2000]) and the triangular ([Revelt and Train, 2000], [Train, 2001]). Covariance between the stochastic elements of ξ are allowed; though applications has been limited to the normal distribution. The unsystematic parts ϵ are IID.

Key issues are to discover and apply the correct shape of the included distributions - including application of a random number generator that does not bias the distribution.

F.3 Drawing a single number

The problem of drawing a single number or a sequence of numbers from a uniform distribution is usually solved by using the random number generator supplied by the computer[2]. Due to the nature of the implementation it is called a pseudo random number generator. The name "pseudo" is chosen because of the relatively high discrepancy, which is a measure of how close to uniform distribution the draws are.

High discrepancy means that the draw is far from uniform and low indicates the draw is close to uniform. This discrepancy can be estimated by selecting a representative subinterval in the range from which the sequence was drawn. The next step is to slide the subinterval through the draw range and for each position where the number of draws in the subinterval change note the count. From the produced set of numbers minimum and maximum is found. The discrepancy can then be described as the difference between these numbers - i.e. if the difference is high so is the discrepancy.

Several approaches to calculating the discrepancy exist. These are interesting if you wish to estimate the discrepancy of a known sequence. In this paper however we consider a special sequence generator, the Halton generator, which by construction yields low discrepancy sequences.

Following the coverage of the pseudo random number draw we cover a widely

[2]More specifically by the operating system

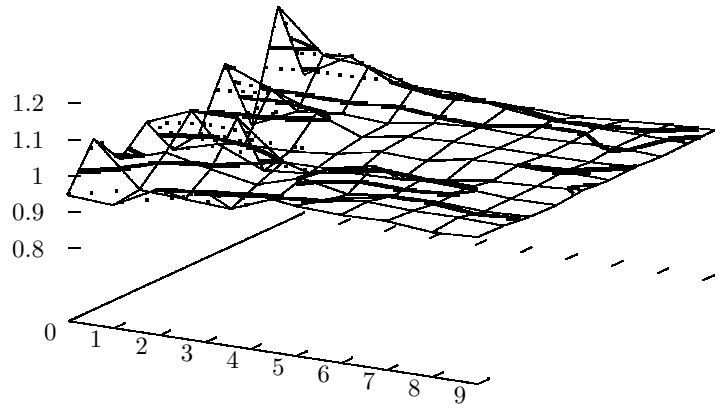


Figure F.1: Visualization: discrepancy of the pseudo random draws

used quasi random number sequence called the Halton sequence and different variations hereof.

F.3.1 Pseudo random numbers

A *pseudo random number* is a general name for numbers generated to simulate uniform distribution draws fast. The crucial characteristic is the speed they are produced at, and not their discrepancy, the deviation from uniformity, as these usually are less important. Usually they are generated as

$$X_{t+1} = mod(aX_t + b, c) \tag{F.3}$$

with proper choices of a, b and c. Further description can be found in eg. [Ross, 1997]. In fact, the discrepancy is the reason why they are insufficient for model estimation.

Figure F.1 shows how significant the discrepancy is. Even if the sequence is extended the discrepancy cannot be neglected, as evident from the figure.

The fallacy is that the numbers are based on a predictable stateful draw that

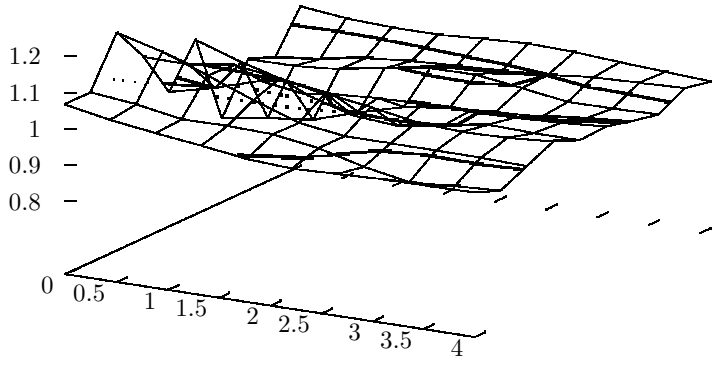


Figure F.2: Visualization: Discrepancy of the entropy based random draw.

sacrifices uniformity and randomness for speed. A newer approach to generating random numbers is based on entropy accummulation and is primarily available under Unix based operating systems. Any bit read from the entropy device is considered to be completely random and highly uniform, which can be translated into: given n-bits from the stream, the number of zeros n_0 equals the number of ones n_1 as $n = n_0 + n_1$ goes to infinity. As with pseudo random numbers the entropy numbers are based on a stateful approach. The major difference is that the state is changed independently of the draw of a number. This, however, is not enough to guarantee sufficiently low discrepancy.

Figure F.2 shows the discrepancy of a 32-bit precision entropy draw. Due to the nature of the entropy generated numbers performance cannot be directly measured, as the time it takes to generate a number is dependent on the entropy of the system. It is thus of no use trying to compare the run times of the two, as the entropy version uses events in the system and hardly any cpu-time as opposed the cpu intense calculations of the standard random draw.

The key point in a sequence generator is not how uniform a single sequence is, but instead the worst discrepancy of all sequences generated by it. The problem in the pseudo draws is that this worst case discrepancy is completely unacceptable. This makes them theoretically useless eventhough practical ex-

amples show that many of the generated sequences have a low discrepancy.

The need for a theoretical guarantee has lead to the invention of sequence generators ensuring a very good worst discrepancy. One, among many of the low discrepancy sequence generators, makes the Halton sequences and we thus call it the Halton generator.

Some of the following have already been covered in great detail in [Bhat, 2001] and [Bhat, 2003], but we include it here as we consider a broader selection of Halton based sequences.

F.3.2 Standard Halton

Rather than ensuring that each number is equally random, the idea is that a sequence of numbers covers the interval uniformingly.

This is where the Halton sequences comes into play. In technical terms is it called a reverse radix-based sequence, where a radical inverse function is used to obtain the point in the interval corresponding to a specific number [Kocis and Whiten, 1997].

Mathematically the n^{th} number in the Halton sequence Φ_b is in [Hess and Polak, 2003] section 3.1 stated as:

$$\Phi_b(n) = \sum_i a_i(n) b^{-i-1}, \text{ where } \sum_{i=0} a_i(n) b^i = n \tag{F.4}$$

Where b is the number generating the sequence and i is the level in the sequence. The problem is the determination of the $a_i(n)$ variables, and the sequences is thus much better described by an example.

The sequence is constructed by continously subdividing the open ended unit interval $[0, 1[$. The first b numbers – the first level – are simply nb^{-1}, dividing the interval into b equally sized subintervals. The next $b^2 - b$ numbers – the second level – divide each subinterval into b new equal sized intervals. After drawing b^2 numbers the interval is divided into b^2 equal sized intervals, which during the next $b^3 - b^2$ (third level) draws is further subdivided into b^3 intervals, etc.

The sequence based on base number 2, is

$$\left\{0, \frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8}, \frac{1}{16} \dots \right\}$$

The sequence of the numbers is such that it is always the first of the widest subintervals that is divided into two by specific proportions based on b.

A standard Halton sequence is far from random, but it is highly uniform as long as the number of draws completes an entire level. In other words the sequence must be of length b^n for $n \in \mathbb{N}$ to satisfy uniformity. If this is not fulfilled the mean of the sequence is be biased towards zero. This is due to the division strategy, where intervals closest to zero are divided first, according to the definition.

This is further strengthened by the first draw of 0, which is never balanced by a draw of 1. This is easily circumvented by simply discarding the 0 and thus using the sequence from the second element, whereby the mean of the sequence $E(\cdot) = \frac{1}{2}$.

F.3.3 Scrambled Halton

The scrambled Halton draw was invented to introduce some randomness into the sequences and, as we will show a little later, alleviate correlation in multi-dimensional draws.

The idea is simply to permute the interval subdivision such that it is not the first of the widest intervals but instead any one of the widest intervals still to be divided in this level, that is split.

The mathematics are given in [Hess and Polak, 2003] section 3.2 as follows:

$$\Phi_b(n) = \sum_i \sigma_b(a_i(n)) b_j^{-i-1}, \text{ where } \sum_{i=0} a_i(n) b^i = n \tag{F.5}$$

The only difference from equation (F.4) is the σ_b function which performs a permutation of all integers from 0 to $b-1$

It is defined above that the permutation function is the same for all levels and given a σ function as $[012] \rightarrow [102]$ the sequence based on 3 becomes:

$$\left\{ \frac{1}{3}, 0, \frac{2}{3}, \frac{4}{9}, \frac{1}{9}, \frac{7}{9}, \frac{5}{9}, \frac{2}{9}, \frac{8}{9}, \frac{13}{27} \dots \right\}$$

Again the sequences are not random, but it is less obvious what the next number is due to the scrambling function σ. It is still highly uniform as the numbers drawn are the same as for the standard Halton, but if a sequence does not complete a level there is no guarantee that the mean $E(\cdot) = \frac{1}{2}$.

The major difference is that the bias caused by incomplete levels is completely dependent on the scrambling function. Choosing a function like $\sigma_5 : [01234] \rightarrow [21304]$ will ensure that the bias is minimized as the interval divisions are performed symmetrically around 0.5.

F.3.4 Leaped Halton

To increase the degree of randomness the leaped Halton draw was invented. It again is based on the standard Halton but instead of using all numbers only every l^{th} number is used. As long as the number l is a prime different from the prime base b it can be shown that the resulting draw is equally uniform. Refer to section 2.4 of [Kocis and Whiten, 1997] for further elaboration on this.

Furthermore it can be shown that the leaping can be considered as a scrambling and pertubation function that rather effectively removes the apparent predictability of the standard and scrambled Halton draws.

The drawback is the possible computational overhead of drawing $n \times l$ elements, instead of just n elements.

The first 10 numbers of the leaped Halton sequence based on $b = 29$ and $l = 409$ are;

0.08561237 0.20570749 0.45782935 0.94415515 0.43052196
0.91684776 0.40321456 0.88954037 0.37590717 0.86223297

with mean ≈ 0.557.

The problem is here that the leap length is constant, which means that while no numbers from the first level are selected in the example above $(29^7 - 29^6)/409 = 40721400$ numbers from level 7 are selected in strictly circular manner. The permutation and pertubation is thus most influencial when the number of draws is relatively small.

F.4 Multidimensional draws

When using multidimensional draws it is important that the different dimensions are uncorrelated - or rather, controlled given that correlation is desired. If and only if they are uncorrelated and uniform, then the draw will be uniform in the s-dimensional unit space $[0, 1[^s$.

F.4.1 Linear draws

The straight forward approach to this is to draw anew for each dimension. As we focus on Halton draws, any of the above can be used. For a 3 dimensional draw we can use three standard Halton sequences each based on a different number b. However, this is not sufficient. Selection of these base numbers are highly important and must be done with great care. The standard Halton sequence is cyclic and inappropriate base numbers introduces high correlation between the dimensions; thus resulting in a non-uniform draw.

Figure F.3 shows correlation between 2 standard Halton sequences.

Even choosing the numbers optimally will not result in a uniform multidimensional draw. The cycles of the individual standard Halton sequences will combine into a supercycle, which will result in correlation.

The first or smallest step to overcome the correlation is to use the *scrambled Halton* instead. This disrupts the cycles of the standard version, but again, dependent on the scrambling function, the cycles may reappear in longer sequences. The correlation is significantly reduced, but has not been removed.

Similarly, using the *leaped Halton* might again reduce the correlation, but the computational overhead can become a drawback.

To define the linear draw more strictly we can define the linear s-dimensional draw $\mathbb{L}^s$, constructed by:

$$\mathbb{L}^s(n) = \{\Sigma_1(n), \Sigma_2(n), \ldots, \Sigma_s(n)\}$$

where $\Sigma_i(n)$ is the n^{th} number in the i^{th} sequence, which can be any sequence from a uniform draw.

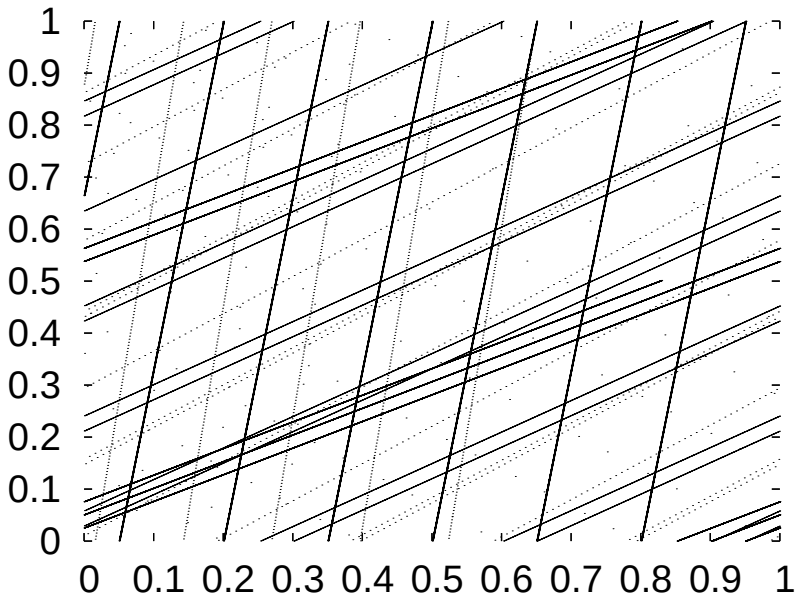


Figure F.3: Visualization: The correlation for generator $b_1 = 11$ and $b_2 = 13$

F.4.2 Shuffled draws

The problem above is that the number drawn for each dimension is drawn from the same sequence in every draw. The idea in shuffled Halton draws is to permute or shuffle the sequences among the dimensions for each draw.

The basic shuffled version proposed lately in [Hess and Polak, 2003] is based on standard Halton sequences shuffled randomly among the dimensions at every draw. This disruption has impact on the correlation as it no longer is exactly two dimensions that are strictly correlated. Instead the correlation might become more or less as the correlated sequences are no longer bound to specific dimensions.

As above, the shuffled draw $\mathbb{S}$ can be described by:

$$\mathbb{S}^s(n) = \sigma_{sn}\left(\Sigma_1(n), \Sigma_2(n), \ldots, \Sigma_s(n)\right)$$

Where σ_{sn} is a function, that dependent on n, permutes the s elements.

A sequence of shufflings could be:
$$\begin{aligned}\sigma_{50}(x_0, x_1, x_2, x_3, x_4) &= (x_2, x_1, x_3, x_0, x_4)\\ \sigma_{51}(x_0, x_1, x_2, x_3, x_4) &= (x_4, x_3, x_1, x_0, x_2)\\ \sigma_{52}(x_0, x_1, x_2, x_3, x_4) &= (x_1, x_4, x_2, x_3, x_0)\end{aligned}$$

F.5 Obtaining the results

The generation time is trivially acquired by examining the per process timing information upon inaugeration and completion of the actual generation.

The correlation between two sequences are found by:

$$r(X,Y) = \frac{Cov(X,Y)}{S(X) \cdot S(Y)}$$

Where $Cov(\cdot,\cdot)$ is the covariance and $S(\cdot)$ is the standard deviation:

$$Cov(X,Y) = E\left[(X - E(X))(Y - E(Y))\right] \text{ and } S(X) = \sqrt{Cov(X,X)}$$

Two one dimensional sequences were defined as correlated if the numerical value of the correlation exceeded $2N^{-1/2}$, that is:

$$|r(X,Y)| > \frac{2}{\sqrt{N}} \Rightarrow \text{Sequences X and Y are correlated}$$

For the multidimensional draws the results are found by first generating a large set of sequences – 100 or more. These were generated from the primes starting from 2. From this initial set of sequences Operations Research was applied to to solve 2 different optimization problems:

1. Find the fixed subset of sequences (10, 20 or more) with the least correlation between all pairs
2. Find the maximum subset of sequences where none of the sequences are correlated with each other

Both problems are very difficult to solve. We have thus set an upper time limit on the applied solver (cplex). The results are the best found solution within the 8 hour time limit we imposed. For almost all of the small (100 and 200) initial sequence sets the results are optimal. As predicted the 500 sequence sets were exponentially harder and the results are the best found when the time limit was exceeded.

F.6 Runtime analysis

As the basics of pseudo random and Halton sequences have been covered, we now turn to an interesting question. Above the draws were seemingly made more and more perfect, but here we try to shed some light on the implementation performance of the different draws.

As shown in [Bhat, 2001] and [Bhat, 2003] applying the Halton sequences through the Quasi-Monte Carlo method to the mixed logit is highly effective. His research however only uses short sequences (25 to 150 numbers) and few dimensions (1 to 10). It is evident that the models applied in transportation research will become increasingly complex and thus require substantially more from the

random sequences used in the estimation of the different models. We thus consider sequences of thousands numbers and up to 500 dimensions as the runtime issue will become even more important. Eventhough the numbers in some cases can be generated a-priori many neglect the use of quasi random numbers due to the extra time it requires. However comparing the runtime analysis in this paper with the effect on MAPE[3] and RMSE[4] in [Bhat, 2001] and [Bhat, 2003] shows that neglecting Halton sequences is a poor disposition.

It must be mentioned that all benchmarks have been optimized and executed on the same uni-processor linux box and the results are thus biased only by the hardware dependent performance of both pseudo random and Halton sequences.

As expected figure F.4 a) shows the the pseudo random draw is approximately three times faster than the standard Halton. Leaping adds roughly 60% to the execution time and finally scrambling doubles it.

Inspecting the multidimensional draws in figure F.4 b) the impression is much less different. The pseudo random draw is fastest, but the difference is not as significant as with the unidimensional draws.

Adding shuffling to the pseudo random numbers is irrelevant which is why pseudo random numbers are left out in figure F.4 c). These compare the impact of adding shuffling to the draw. The "full" refers to full permutation which is based on a slower, but theoretically sound, stateful approach consuming $O(b)$ space where as the "random" refers to a generic permutation based on projection requiring $O(1)$ space.

It can be concluded that the performance impact of the full permutation as opposed to generic permutation, is negligible and that the shuffling adds approximately 10 percent to the generation time of the fully shuffled and scrambled sequence.

[3]Mean Absolute Percentage Error (MAPE)
[4]Root Mean Square Error (RMSE)

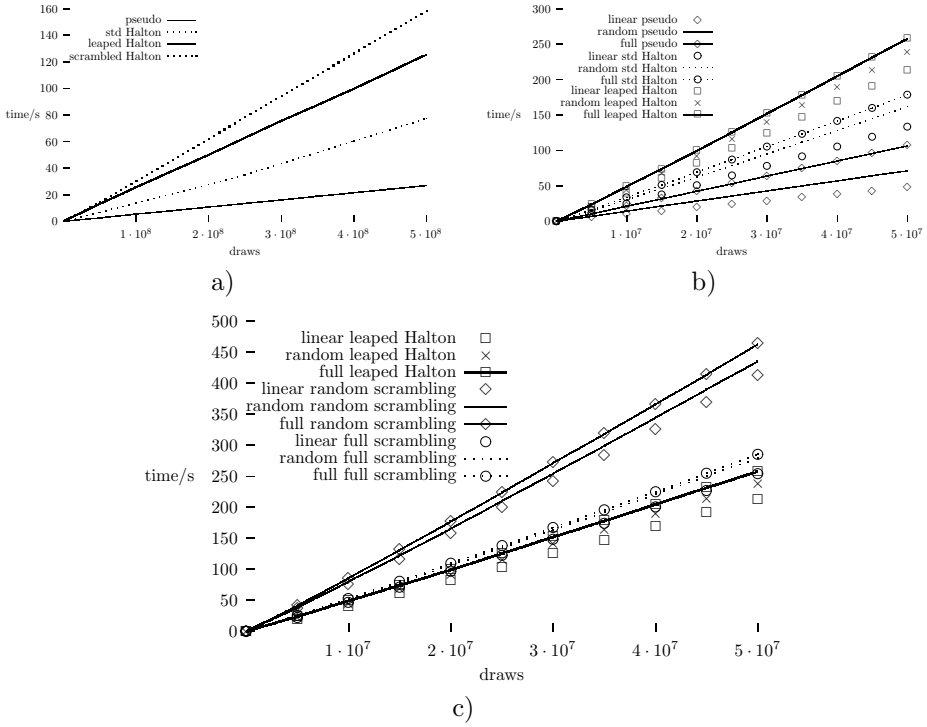


Figure F.4: Execution times for a) 1-D draws and b) 5-D draws using different shufflings. c) compares the linear leaped Halton with different shuffling/scrambling combinations (eg full random scrambled means fully shuffled randomly scrambled Halton)

F.7 Correlation

To evaluate the influence of the shuffling, leaping or scrambling on a multidimensional draw several issues have to be considered. First we consider the influence of leaping and scrambling compared to the length of the sequence. After this we will attend to the multidimensional draws and shuffling.

Figure F.5 indicates that the correlation generally is reduced as the sequences are extended. The problem is that the reduction is less than n^{-2}, which indicates that the significance of the correlation increases as the sequences get longer.

There is one very important exception from the above rule; which is the illustrations in rows two and four where the leaped sequences have been used. At a specific point the correlation suddenly starts increasing. This is because the sequences were leaped with the same length and that this leap length becomes dominant in the number generation. This increase happens much earlier for the highly correlated pairs than for the less correlated pairs.

The small periodic "bubbles" on the standard graphs are caused by the cyclic nature of the Halton sequence.

The graphs show that the low correlated pair is relatively unaffected by choice of generation whereas the highly correlated pair becomes better when using leaping, scrambling or both.

As leaping has the possible correlation increase we would recommend using the scrambled Halton sequences.

Figure F.6 and figure F.8 shows the main findings of the multidimensional tests. These were found by solving the optimization problems described in section F.5

Several guidelines can be deducted by inspecting the numbers.

First, increasing the number of initially generated draws does not reduce the correlation of the selected set. Which means that the least correlated pair are found among the smaller primes. Second, scrambling is the most effective measure to eliminate correlation in larger subset selections.

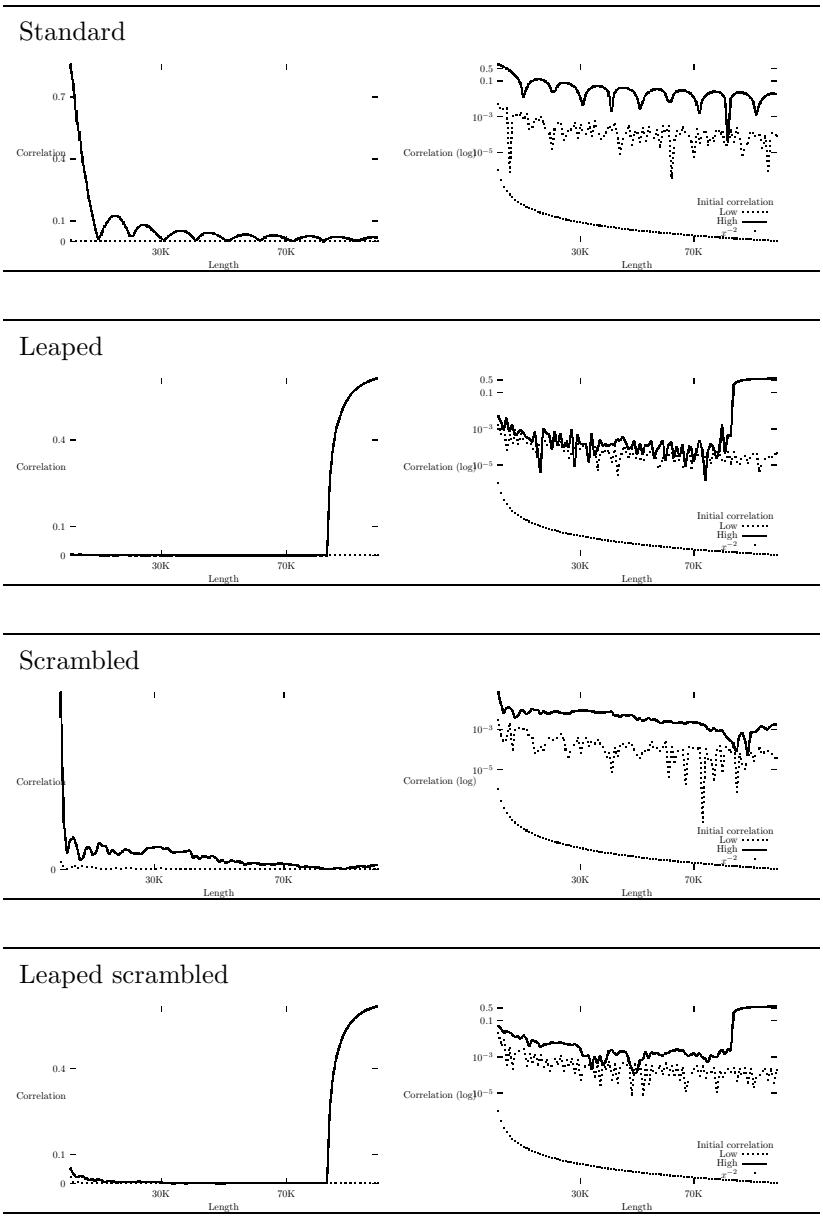


Figure F.5: Length and correlation

Dimensions: 200	Selection			
draws: 1000	10	20	30	50
halton-linear	0.00617	∼ 0.14617	∼ 0.79661	0.96039
halton-shuffled	∼ 0.15547	∼ 0.19574	∼ 0.22030	∼ 0.28451
leaped-linear	∼ 0.00931	∼ 0.04764	∼ 0.30843	0.94363
leaped-shuffled	∼ 0.06414	∼ 0.09291	∼ 0.11466	∼ 0.14040
scrambled-linear	∼ 0.01485	∼ 0.04588	∼ 0.07084	∼ 0.09108
scrambled-shuffled	∼ 0.02432	∼ 0.04563	∼ 0.06466	∼ 0.10614

This table summarizes the result of optimization problem 1 with initial set size 200 and sequence length of 1000. It can be seen that the lowest pairwise correlation is found with standard halton sequences with out shuffling for the small set. In all other cases scrambled halton sequences yeild the best result. This was also the case with all other initial set sizes 100, 300, 500 and 1000.

Figure F.6: Minimal correlation with preset number of generated sequences

Selection: 50	Dimensions				
draws: 1000	100	200	300	500	1000
halton-linear	∼ 0.53868	0.96039	1.00000	1.00000	1.00000
halton-shuffled	∼ 0.12817	∼ 0.28451	∼ 0.32453	∼ 0.31940	0.26208
leaped-linear	∼ 0.30618	0.94363	1.00000	∼ 0.99348	1.00000
leaped-shuffled	∼ 0.08794	∼ 0.14040	∼ 0.18867	∼ 0.20099	0.31337
scrambled-linear	∼ 0.06287	∼ 0.09108	∼ 0.10321	∼ 0.10040	0.10532
scrambled-shuffled	∼ 0.06814	∼ 0.10614	∼ 0.09947	∼ 0.10114	0.14266

This table shows the same results as figure F.6, but the influence of the initial set size is clearer here. The linear multidimensional draws shows that the solver is having an increasingly hard time as the initial set size increases. This can be seen from the fact that the correlation rises with the size of the initial set even though the initial set of 200 contains the sequences in the size 100 initial set.

The shuffled draws however shows that increasing the initial size tends to increase the found correlation. This could be caused by the fact that the halton sequences of the larger size initial sets are more correlated than the smaller initial sets.

Figure F.7: Minimal correlation with preset selection size

Maximal uncorrelated	Dimensions				
subset (draws: 1000)	100	200	300	500	1000
halton-linear	25	26	26	26	26
halton-shuffled	7	2	1	1	2
leaped-linear	30	31	32	32	33
leaped-shuffled	23	12	≥ 6	3	2
scrambled-linear	49	67	≥ 74	≥ 84	≥ 93
scrambled-shuffled	42	≥ 62	≥ 71	≥ 82	≥ 90

This table shows the results from optimization problem 2. It can easily be seen that shuffling is deteriorating and that scrambling is significantly better if getting the maximum uncorrelated set is most important. The results where $\geq$ # is used means that the solver reached the time limit before optimality, but had found a feasible solution with # uncorrelated sequences. As the solver is highly efficient and quickly gets close to optimality it is unlikely that sets with significantly more sequences exists in the initial set.

Inspecting the subset sizes for the scrambled linear generator shows that it will be extremely demanding (huge initial set and very long computation time) to find a subset of more than 100 uncorrelated sequences when using halton sequences.

Figure F.8: Maximal sets

A third observation, that is less obvious, is that the introduction of shuffling has negative impact on small sets, while yielding lower correlation on large sets. The explanation for this is that the shuffling evens the correlation among the generated sequences. In other words; it sacrifices the low correlation pairs in order to get better high correlation pairs.

Furthermore it can be concluded that using both shuffling and scrambling seems needless as the scrambling leaves little room for improvement by the shuffling.

It can be concluded that standard Halton sequences are best for draws with less than approximately 15 dimensions. Further sequence inspection indicated that selecting bases among the first 40 primes is preferable. From 15 to around 25 dimensions and limited length leaped Halton sequences should be used. For draws of larger dimension scrambling is the most effective measure for reducing the correlation.

If uniform correlation is important, shuffling can be used to bring the pairwise correlation closer to average.

Appendices F.9, F.10, F.11 and F.12 gives visual examples on a highly correlated set.

F.8 Conclusion

Building from the standard Halton draw we have now been through the two other single dimension number sequences: leaped and scrambled and we have discussed the multidimensional shuffling.

Several combinations of these have been compared and it can be concluded that the usage of the multidimentional sequence must be considered when deciding how to generate the numbers.

As basic sequence the scrambled Halton sequence provide the least interdimensional correlation. Leaping is also a possibility, but the leaping distances must be chosen carefully as they introduce correlation in long sequences.

Shuffling does not guarantee a reduction in the correlation of a draw, but instead it evens the correlation such that every pair of sequences in a multidimensional draw becomes closer to equally correlated.

Considering a single measure to avoid correlation it can be concluded scrambling gives lowest correlation at the cost of slightly longer generation time.

This paper has demonstrated that there are immediately unobservable – though detectable, differences between the performance of random numbers, contingent on their generation.

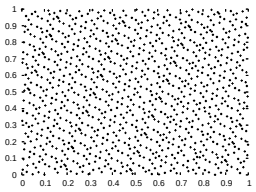
The differences in the uniformity and bias will inevitably impact the modelling of any model, where random numbers are incorporated by use of simulation in the estimations as eg. the Mixed Logit estimation.

Coefficients estimated and even distributions thought valid for a given model may be impacted by the use of different random number generators. The impact will be via the calibration, hence the estimation provided an iterative method is applied for the estimation of the model.

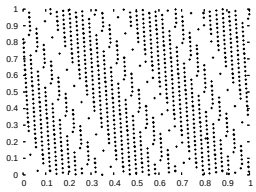
Therefore, use of method for random number generation should actively be considered and at least, be documented along with other model formulation considerations.

F.9 The pseudorandom draw

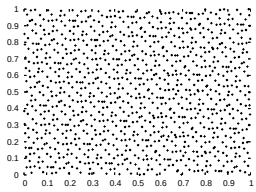
Visualization of the linear combination where the sequences are statically assigned to dimensions.



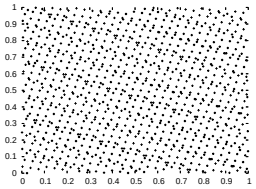
1 and 2



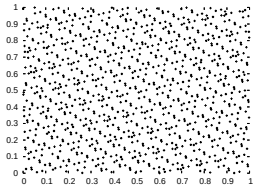
1 and 3



1 and 4



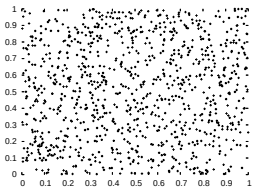
1 and 5



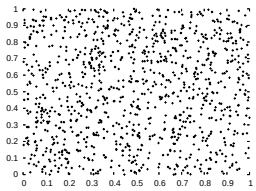
2 and 3

1, 2 and 3

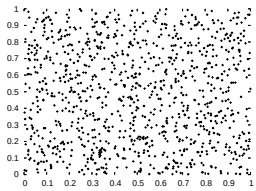
The shuffled approach where the sequences are shuffled at each draw.



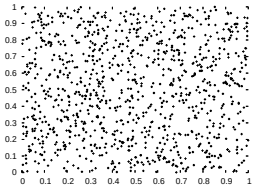
1 and 2



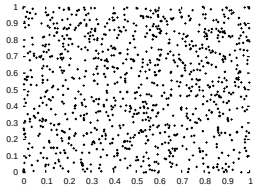
1 and 3



1 and 4



1 and 5

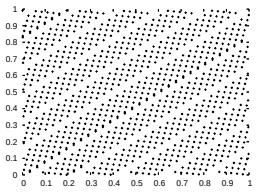


2 and 3

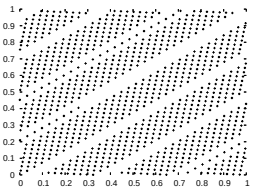
1, 2 and 3

F.10 The standard Halton draw

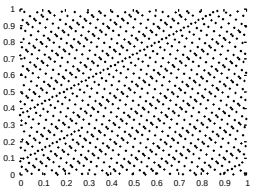
Visualization of the linear combination where the sequences are statically assigned to dimensions.



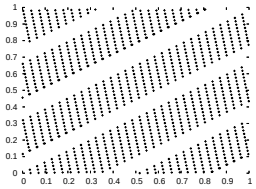
Φ_{37} and Φ_{41}



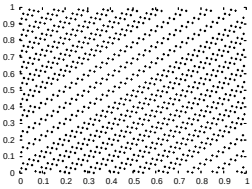
Φ_{37} and Φ_{43}



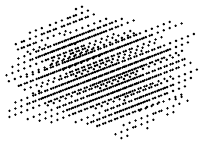
Φ_{37} and Φ_{47}



Φ_{37} and Φ_{53}

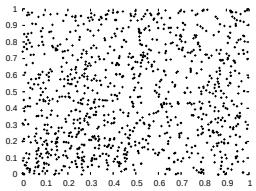


Φ_{41} and Φ_{43}

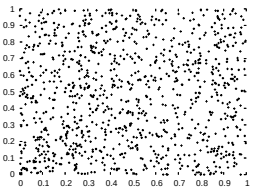


Φ_{37}, Φ_{41} and Φ_{43}

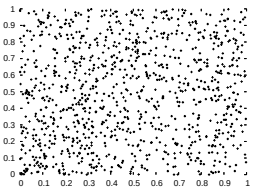
The shuffled approach where the sequences are shuffled at each draw.



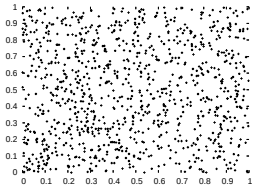
Φ_{37} and Φ_{41}



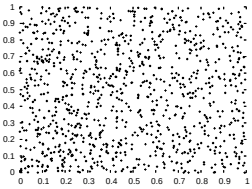
Φ_{37} and Φ_{43}



Φ_{37} and Φ_{47}



Φ_{37} and Φ_{53}

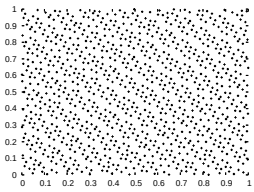


Φ_{41} and Φ_{43}

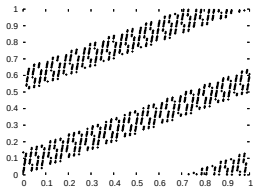
Φ_{37}, Φ_{41} and Φ_{43}

F.11 The leaped Halton draw

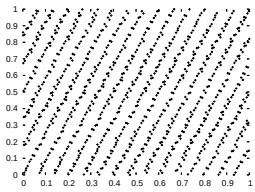
Visualization of the linear combination where the sequences are statically assigned to dimensions.



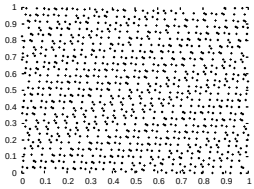
$\Phi_{73^\frown 79}$ and $\Phi_{83^\frown 89}$



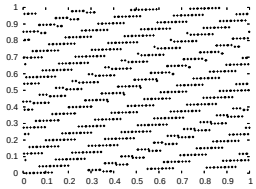
$\Phi_{73^\frown 79}$ and $\Phi_{97^\frown 101}$



$\Phi_{73^\frown 79}$ and $\Phi_{103^\frown 107}$



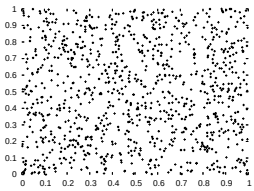
$\Phi_{73^\frown 79}$ and $\Phi_{109^\frown 113}$



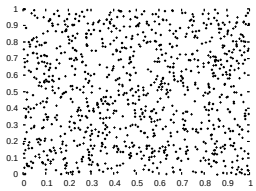
$\Phi_{83^\frown 89}$ and $\Phi_{97^\frown 101}$

$\Phi_{73^\frown 79}$, $\Phi_{83^\frown 89}$ and $\Phi_{97^\frown 101}$

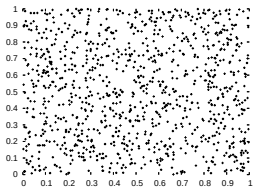
The shuffled approach where the sequences are shuffled at each draw.



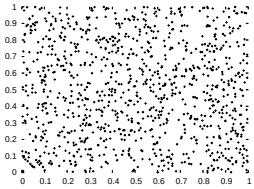
$\Phi_{73^\frown 79}$ and $\Phi_{83^\frown 89}$



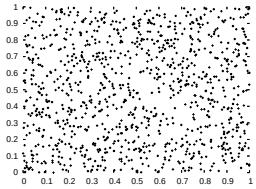
$\Phi_{73^\frown 79}$ and $\Phi_{97^\frown 101}$



$\Phi_{73^\frown 79}$ and $\Phi_{103^\frown 107}$



$\Phi_{73^\frown 79}$ and $\Phi_{109^\frown 113}$

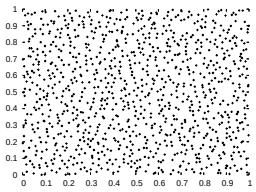


$\Phi_{83^\frown 89}$ and $\Phi_{97^\frown 101}$

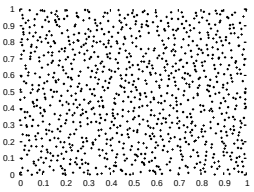
$\Phi_{73^\frown 79}$, $\Phi_{83^\frown 89}$ and $\Phi_{97^\frown 101}$

F.12 The scrambled Halton draw

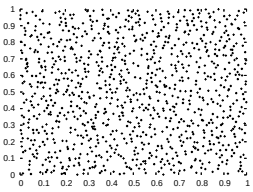
Visualization of the linear combination where the sequences are statically assigned to dimensions.



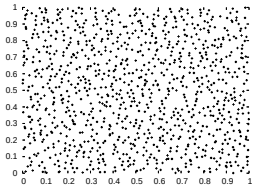
Φ_{37} and Φ_{41}



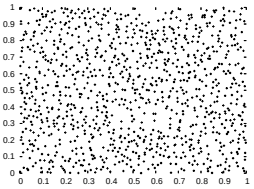
Φ_{37} and Φ_{43}



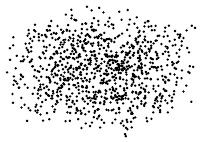
Φ_{37} and Φ_{47}



Φ_{37} and Φ_{53}

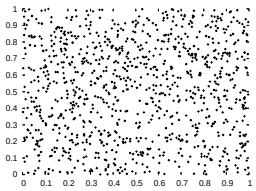


Φ_{41} and Φ_{43}

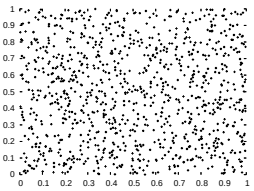


Φ_{37}, Φ_{41} and Φ_{43}

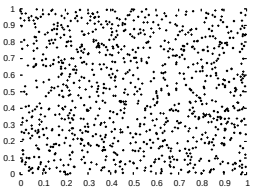
The shuffled approach where the sequences are shuffled at each draw.



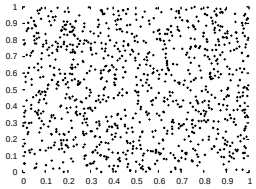
Φ_{37} and Φ_{41}



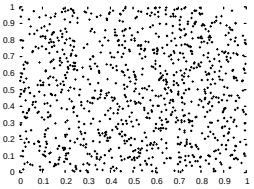
Φ_{37} and Φ_{43}



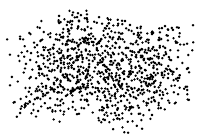
Φ_{37} and Φ_{47}



Φ_{37} and Φ_{53}



Φ_{41} and Φ_{43}



Φ_{37}, Φ_{41} and Φ_{43}

Bibliography

[Ben-Akiva et al., 1993] Ben-Akiva, M., Bolduc, D., and Bradley, M. (1993). Estimation of travel choice models with random distributed values of time. *Transportation Research Record*, 1413:88–97.

[Ben-Akiva et al., 2001] Ben-Akiva, M., Bolduc, D., and Walker, J. L. (2001). Specification, identification & estimation of the logit kernel (or continuous mixed logit) model. Draft presented at 5th tri-annual Invitational Choice Symposium Hosted by UC Berkeley, June 1-5, 2001.

[Ben-Akiva et al., 2002] Ben-Akiva, M., Mcfadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Bolduc, D., Boersch-Supan, A., Brownstone, D., Bunch, D., Daly, A., De Palma, A., Gopinath, D., Karlstrom, A., and Munizaga, M. (2002). Hybrid choice models: Progress and challenges. *Marketing Letters*, 13 (3):165–175.

[Bhat, 2003] Bhat, C. (2003). Simulation estimation of mixed discrete choice models using randomized and scrambled halton sequences. *Transportation Research Part B: Methodological*, 37(9):837–855.

[Bhat, 2001] Bhat, C. R. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological*, 35(7):677–693.

[Bierlaire, 2002] Bierlaire, M. (2002). Biogeme website. http://roso.epfl.ch/biogeme.

[Hess and Polak, 2003] Hess, S. and Polak, J. W. (2003). On the performance of the shuffled halton sequence in the estiamtion od discrete choice models. Technical report, Association for European Transport 2003.

[Kocis and Whiten, 1997] Kocis, L. and Whiten, W. J. (1997). Computational investigations of low-discrepancy sequences. *ACM Transactions on Mathematical Software*, 23(2):266–294.

[McFadden and Train, 2000] McFadden, D. and Train, K. (2000). Mixed mnl models for discrete response. *Journal of Applied Econometrics*, 15(5):447–70.

[Nunes et al., 2001] Nunes, L. C., Cunha-e Sa, M. A., Ducla-Soares, M. M., Rosado, M. A., and Day, B. H. (2001). Identifying non-consistent choice behavior in recreation demand models. *Economics Letters*, 72(3):403–410.

[Revelt and Train, 2000] Revelt, D. and Train, K. E. (2000). Customer-specific taste parameters and mixed logit. Working Paper 00-274, Department of Economics, University of California, Berkeley.

[Ross, 1997] Ross, S. M. (1997). *Simulation*. Academic Press, Burlington, MA, USA, 2nd edition.

[Sørensen, 2003a] Sørensen, M. V. (2003a). *Discrete Choice Models. Estimation of passenger traffic.* PhD thesis, Danish Technical University, DK-2800 Kgs. Lyngby, Denmark.

[Sørensen, 2003b] Sørensen, M. V. (2003b). Msl for mixed logit model estimation - on shape of distributions. In *Proceedings of European Transport Conference, Strasbourg, France*. CD-rom.

[Sørensen, 2004] Sørensen, M. V. (2004). Impact of a priori distributions on mixed logit model estimation. tests on synthetic data. In *Proceedings of World Conference on Transport Research, Istanbul , Turkey*.

[Train, 1998] Train, K. E. (1998). Recreation demand models with taste differences over people. *Land Economics*, 74(2):230–239.

[Train, 2001] Train, K. E. (2001). A comparison of hierarchical bayes and maximum simulated likelihood for mixed logit. Working paper, Department of Economics, University of California, Berkeley.

[Train, 2002] Train, K. E. (2002). Website with gauss shareware code. http://elsa.berkeley.edu/~train/software.html.

Appendix G

Acronyms

ABIT Agent Based Individual Traffic guidance

AC Ambient Computing

ATIS Advanced Traffic Information System

ATMS Advanced Traffic Management Systems

BFGS Broyden-Fletcher-Goldfarb-Shanno

CARG Consistent Anticipatory Route Guidance

CIS Central Information System

CP Complementarity Problems

DFP Davidon-Fletcher-Powell

DM Disruption Management

DRIVE General European Road Data Information Exchange Network

DTA Dynamic Traffic Assignment

DTS Dynamic Traffic Signs

EDI Electronic Data Interchange

EFS Efficient Forward Star

ETD Expected Traveling Distance

ETT Expected Traveling Time

FD Fluid Dynamics

FW Frank-Wolfe method

GAF gap acceptancy function

GIS Geographic Information Systems

GPS Global Positioning System

GPRS Global Packet Radio Service

ICT Information and Communications Technoligies

INFORMS Institute for Operations Research and the Management Sciences

IP Integer Programming

ITS Intelligent Traffic Systems

IVHS Intelligent Vehicle-Highway Systems

ISTEA Intermodal Surface Transportation Efficiency Act

LoI Level of Influence

LIWAS Life Warning System

LP Linear Programming

MCP Mixed Complementarity Problems

MIP Mixed Integer Programming

MP Mathematical Programming

MSA Method of Successive Averages

NLP Non-Linear Programming

OBA Origin-Based Algorithm

OR Operations Research

PC Pervasive Computing

PI Pervasive Intelligence

PNM Pseudo-Newton Method

PTI Pervasive Traffic Intelligence

QNM Quasi-Newton Method

QP Quadratic Programming

RAF Royal Air Force

RDS Radio Data System

TAP Traffic Assignment Problem

TMC Traffic Message Channel

TR Transportation Research

TS Traffic Science

VICS Vehicle Information and Communication System

VIP Variational Inequality Problems

VMS Variable Message Signs

WLAN Wireless Local Area Network